

---

# Table of Contents

Introduction	1.1
Spark Structured Streaming and Streaming Queries	1.2
Batch Processing Time	1.2.1
Internals of Streaming Queries	1.3

## Streaming Join

Streaming Join	2.1
StateStoreAwareZipPartitionsRDD	2.2
SymmetricHashJoinStateManager	2.3
StateStoreHandler	2.3.1
KeyToNumValuesStore	2.3.2
KeyWithIndexToValueStore	2.3.3
OneSideHashJoiner	2.4
JoinStateWatermarkPredicates	2.5
JoinStateWatermarkPredicate	2.5.1
StateStoreAwareZipPartitionsHelper	2.6
StreamingSymmetricHashJoinHelper	2.7
StreamingJoinHelper	2.8

## Extending Structured Streaming with New Data Sources

Extending Structured Streaming with New Data Sources	3.1
BaseStreamingSource	3.2
BaseStreamingSink	3.3
StreamWriteSupport	3.4
StreamWriter	3.5
DataSource	3.6

---

# Demos

Demos	4.1
Internals of FlatMapGroupsWithStateExec Physical Operator	4.2
Arbitrary Stateful Streaming Aggregation with KeyValueGroupedDataset.flatMapGroupsWithState Operator	4.3
Exploring Checkpointed State	4.4
Streaming Watermark with Aggregation in Append Output Mode	4.5
Streaming Query for Running Counts (Socket Source and Complete Output Mode)	4.6
Streaming Aggregation with Kafka Data Source	4.7
groupByKey Streaming Aggregation in Update Mode	4.8
StateStoreSaveExec with Complete Output Mode	4.9
StateStoreSaveExec with Update Output Mode	4.10
Developing Custom Streaming Sink (and Monitoring SQL Queries in web UI)	4.11
current_timestamp Function For Processing Time in Streaming Queries	4.12
Using StreamingQueryManager for Query Termination Management	4.13

# Streaming Aggregation

Streaming Aggregation	5.1
StateStoreRDD	5.2
StateStoreOps	5.2.1
StreamingAggregationStateManager	5.3
StreamingAggregationStateManagerBaseImpl	5.3.1
StreamingAggregationStateManagerImplV1	5.3.2
StreamingAggregationStateManagerImplV2	5.3.3

# Stateful Stream Processing

Stateful Stream Processing	6.1
Streaming Watermark	6.2
Streaming Deduplication	6.3
Streaming Limit	6.4
StateStore	6.5

---

StateStoreId	6.5.1
HDFSBackedStateStore	6.5.2
StateStoreProvider	6.6
StateStoreProviderId	6.6.1
HDFSBackedStateStoreProvider	6.6.2
StateStoreCoordinator	6.7
StateStoreCoordinatorRef	6.7.1
WatermarkSupport	6.8
StatefulOperator	6.9
StateStoreReader	6.9.1
StateStoreWriter	6.9.2
StatefulOperatorStateInfo	6.10
StateStoreMetrics	6.11
StateStoreCustomMetric	6.12
StateStoreUpdater	6.13
EventTimeStatsAccum	6.14
StateStoreConf	6.15

## Arbitrary Stateful Streaming Aggregation

Arbitrary Stateful Streaming Aggregation	7.1
GroupState	7.2
GroupStateImpl	7.2.1
GroupStateTimeout	7.3
StateManager	7.4
StateManagerImplV2	7.4.1
StateManagerImplBase	7.4.2
StateManagerImplV1	7.4.3
FlatMapGroupsWithStateExecHelper Helper Class	7.5
InputProcessor Helper Class of FlatMapGroupsWithStateExec Physical Operator	7.6

## Developing Streaming Applications

DataStreamReader	8.1
------------------	-----

---

DataStreamWriter	8.2
OutputMode	8.2.1
Trigger	8.2.2
StreamingQuery	8.3
Streaming Operators	8.4
dropDuplicates Operator	8.4.1
explain Operator	8.4.2
groupBy Operator	8.4.3
groupByKey Operator	8.4.4
withWatermark Operator	8.4.5
window Function	8.5
KeyValueGroupedDataset	8.6
mapGroupsWithState Operator	8.6.1
flatMapGroupsWithState Operator	8.6.2
StreamingQueryManager	8.7
SQLConf	8.8
Configuration Properties	8.9

## Monitoring of Streaming Query Execution

StreamingQueryListener	9.1
ProgressReporter	9.2
StreamingQueryProgress	9.3
ExecutionStats	9.3.1
SourceProgress	9.3.2
SinkProgress	9.3.3
StreamingQueryStatus	9.4
MetricsReporter	9.5
Web UI	9.6
Logging	9.7

## File-Based Data Source

FileStreamSource	10.1
------------------	------

---

<a href="#">FileStreamSink</a>	10.2
<a href="#">FileStreamSinkLog</a>	10.3
<a href="#">SinkFileStatus</a>	10.4
<a href="#">ManifestFileCommitProtocol</a>	10.5
<a href="#">MetadataLogFileIndex</a>	10.6

---

## Kafka Data Source

<a href="#">Kafka Data Source</a>	11.1
<a href="#">KafkaSourceProvider</a>	11.2
<a href="#">KafkaSource</a>	11.3
<a href="#">KafkaRelation</a>	11.4
<a href="#">KafkaSourceRDD</a>	11.5
<a href="#">CachedKafkaConsumer</a>	11.6
<a href="#">KafkaSourceOffset</a>	11.7
<a href="#">KafkaOffsetReader</a>	11.8
<a href="#">ConsumerStrategy</a>	11.9
<a href="#">KafkaSink</a>	11.10
<a href="#">KafkaOffsetRangeLimit</a>	11.11
<a href="#">KafkaDataConsumer</a>	11.12
<a href="#">KafkaMicroBatchReader</a>	11.13
<a href="#">KafkaOffsetRangeCalculator</a>	11.13.1
<a href="#">KafkaMicroBatchInputPartition</a>	11.13.2
<a href="#">KafkaMicroBatchInputPartitionReader</a>	11.13.3
<a href="#">KafkaSourceInitialOffsetWriter</a>	11.13.4
<a href="#">KafkaContinuousReader</a>	11.14
<a href="#">KafkaContinuousInputPartition</a>	11.14.1

---

## Text Socket Data Source

<a href="#">TextSocketSourceProvider</a>	12.1
<a href="#">TextSocketSource</a>	12.2

---

---

## Rate Data Source

<a href="#">RateSourceProvider</a>	13.1
<a href="#">RateStreamSource</a>	13.2
<a href="#">RateStreamMicroBatchReader</a>	13.3

---

## Console Data Sink

<a href="#">ConsoleSinkProvider</a>	14.1
<a href="#">ConsoleWriter</a>	14.2

---

## Foreach Data Sink

<a href="#">ForeachWriterProvider</a>	15.1
<a href="#">ForeachWriter</a>	15.2
<a href="#">ForeachSink</a>	15.3

---

## ForeachBatch Data Sink

<a href="#">ForeachBatchSink</a>	16.1
----------------------------------	------

---

## Memory Data Source

<a href="#">Memory Data Source</a>	17.1
<a href="#">MemoryStream</a>	17.2
<a href="#">ContinuousMemoryStream</a>	17.3
<a href="#">MemorySink</a>	17.4
<a href="#">MemorySinkV2</a>	17.5
<a href="#">MemoryStreamWriter</a>	17.5.1
<a href="#">MemoryStreamBase</a>	17.6
<a href="#">MemorySinkBase</a>	17.7

---

---

# Offsets and Metadata Checkpointing (Fault-Tolerance and Reliability)

Offsets and Metadata Checkpointing	18.1
MetadataLog	18.2
HDFSMetadataLog	18.3
CommitLog	18.4
CommitMetadata	18.4.1
OffsetSeqLog	18.5
OffsetSeq	18.5.1
CompactibleFileStreamLog	18.6
FileStreamSourceLog	18.6.1
OffsetSeqMetadata	18.7
CheckpointFileManager	18.8
FileContextBasedCheckpointFileManager	18.8.1
FileSystemBasedCheckpointFileManager	18.8.2
Offset	18.9
StreamProgress	18.10

# Micro-Batch Stream Processing (Structured Streaming V1)

Micro-Batch Stream Processing	19.1
MicroBatchExecution	19.2
MicroBatchWriter	19.2.1
MicroBatchReadSupport	19.3
MicroBatchReader	19.3.1
WatermarkTracker	19.4
Source	19.5
StreamSourceProvider	19.5.1
Sink	19.6
StreamSinkProvider	19.6.1

---

# Continuous Stream Processing (Structured Streaming V2)

Continuous Stream Processing	20.1
ContinuousExecution	20.2
ContinuousReadSupport Contract	20.3
ContinuousReader Contract	20.4
RateStreamContinuousReader	20.5
EpochCoordinator RPC Endpoint	20.6
EpochCoordinatorRef	20.6.1
EpochTracker	20.6.2
ContinuousQueuedDataReader	20.7
DataReaderThread	20.7.1
EpochMarkerGenerator	20.7.2
PartitionOffset	20.8
ContinuousExecutionRelation Leaf Logical Operator	20.9
WriteToContinuousDataSource Unary Logical Operator	20.10
WriteToContinuousDataSourceExec Unary Physical Operator	20.11
ContinuousWriteRDD	20.11.1
ContinuousDataSourceRDD	20.12

## Query Planning and Execution

StreamExecution	21.1
StreamingQueryWrapper	21.1.1
TriggerExecutor	21.2
IncrementalExecution	21.3
StreamingQueryListenerBus	21.4
StreamMetadata	21.5

## Logical Operators

EventTimeWatermark Unary Logical Operator	22.1
FlatMapGroupsWithState Unary Logical Operator	22.2

---

Deduplicate Unary Logical Operator	22.3
MemoryPlan Logical Query Plan	22.4
StreamingRelation Leaf Logical Operator for Streaming Source	22.5
StreamingRelationV2 Leaf Logical Operator	22.6
StreamingExecutionRelation Leaf Logical Operator for Streaming Source At Execution	
	22.7

## Physical Operators

EventTimeWatermarkExec	23.1
FlatMapGroupsWithStateExec	23.2
StateStoreRestoreExec	23.3
StateStoreSaveExec	23.4
StreamingDeduplicateExec	23.5
StreamingGlobalLimitExec	23.6
StreamingRelationExec	23.7
StreamingSymmetricHashJoinExec	23.8

## Execution Planning Strategies

FlatMapGroupsWithStateStrategy	24.1
StatefulAggregationStrategy	24.2
StreamingDeduplicationStrategy	24.3
StreamingGlobalLimitStrategy	24.4
StreamingJoinStrategy	24.5
StreamingRelationStrategy	24.6

## Varia

UnsupportedOperationChecker	25.1
-----------------------------	------

# The Internals of Spark Structured Streaming (Apache Spark 2.4.4)

Welcome to **The Internals of Spark Structured Streaming** gitbook! I'm very excited to have you here and hope you will enjoy exploring the internals of Spark Structured Streaming as much as I have.

I write to discover what I know.

— Flannery O'Connor

I'm [Jacek Laskowski](#), a freelance IT consultant, software engineer and technical instructor specializing in [Apache Spark](#), [Apache Kafka](#) and [Kafka Streams](#) (with [Scala](#) and [sbt](#)).

I offer software development and consultancy services with hands-on in-depth workshops and mentoring. Reach out to me at [jacek@japila.pl](mailto:jacek@japila.pl) or [@jaceklaskowski](https://twitter.com/jaceklaskowski) to discuss opportunities.

Consider joining me at [Warsaw Scala Enthusiasts](#) and [Warsaw Spark](#) meetups in Warsaw, Poland.

Tip

I'm also writing other books in the "The Internals of" series about [Apache Spark](#), [Spark SQL](#), [Apache Kafka](#), and [Kafka Streams](#).

Expect text and code snippets from a variety of public sources. Attribution follows.

Now, let me introduce you to [Spark Structured Streaming and Streaming Queries](#).

# Spark Structured Streaming and Streaming Queries

**Spark Structured Streaming** (aka *Structured Streaming* or *Spark Streams*) is the module of Apache Spark for stream processing using **streaming queries**.

Streaming queries can be expressed using a [high-level declarative streaming API](#) (*Dataset API*) or good ol' SQL (*SQL over stream / streaming SQL*). The declarative streaming Dataset API and SQL are executed on the underlying highly-optimized Spark SQL engine.

The semantics of the Structured Streaming model is as follows (see the article [Structured Streaming In Apache Spark](#)):

At any time, the output of a continuous application is equivalent to executing a batch job on a prefix of the data.

Note	As of Spark 2.2.0, Structured Streaming has been marked stable and ready for production use. With that the other older streaming module <b>Spark Streaming</b> is considered obsolete and not used for developing new streaming applications with Apache Spark.
------	---

Spark Structured Streaming comes with two [stream execution engines](#) for executing streaming queries:

- [MicroBatchExecution](#) for [Micro-Batch Stream Processing](#)
- [ContinuousExecution](#) for [Continuous Stream Processing](#)

The goal of Spark Structured Streaming is to unify streaming, interactive, and batch queries over structured datasets for developing end-to-end stream processing applications dubbed **continuous applications** using Spark SQL's Datasets API with additional support for the following features:

- [Streaming Aggregation](#)
- [Streaming Join](#)
- [Streaming Watermark](#)
- [Arbitrary Stateful Streaming Aggregation](#)
- [Stateful Stream Processing](#)

In Structured Streaming, Spark developers describe custom streaming computations in the same way as with Spark SQL. Internally, Structured Streaming applies the user-defined structured query to the continuously and indefinitely arriving data to analyze real-time streaming data.

Structured Streaming introduces the concept of **streaming datasets** that are *infinite datasets* with primitives like input **streaming data sources** and output **streaming data sinks**.

A `dataset` is **streaming** when its logical plan is streaming.

```
val batchQuery = spark.  
  read. // <-- batch non-streaming query  
  csv("sales")  
  
assert(batchQuery.isStreaming == false)  
  
val streamingQuery = spark.  
  readStream. // <-- streaming query  
  format("rate").  
  load  
  
assert(streamingQuery.isStreaming)
```

Tip

Read up on Spark SQL, Datasets and logical plans in [The Internals of Spark SQL book](#).

Structured Streaming models a stream of data as an infinite (and hence continuous) table that could be changed every streaming batch.

You can specify **output mode** of a streaming dataset which is what gets written to a streaming sink (i.e. the infinite result table) when there is a new data available.

Streaming Datasets use **streaming query plans** (as opposed to regular batch Datasets that are based on batch query plans).

From this perspective, batch queries can be considered streaming Datasets executed once only (and is why some batch queries, e.g. [KafkaSource](#), can easily work in batch mode).

Note

```
val batchQuery = spark.read.format("rate").load  
  
assert(batchQuery.isStreaming == false)  
  
val streamingQuery = spark.readStream.format("rate").load  
  
assert(streamingQuery.isStreaming)
```

With Structured Streaming, Spark 2 aims at simplifying **streaming analytics** with little to no need to reason about effective data streaming (trying to hide the unnecessary complexity in your streaming analytics architectures).

Structured streaming is defined by the following data abstractions in `org.apache.spark.sql.streaming` package:

- [StreamingQuery](#)
- [Streaming Source](#)
- [Streaming Sink](#)
- [StreamingQueryManager](#)

Structured Streaming follows micro-batch model and periodically fetches data from the data source (and uses the `DataFrame` data abstraction to represent the fetched data for a certain batch).

With Datasets as Spark SQL's view of structured data, structured streaming checks input sources for new data every [trigger](#) (time) and executes the (continuous) queries.

Note	The feature has also been called <b>Streaming Spark SQL Query</b> , <b>Streaming DataFrames</b> , <b>Continuous DataFrame</b> or <b>Continuous Query</b> . There have been lots of names before the Spark project settled on Structured Streaming.
------	--

## Further Reading Or Watching

- [SPARK-8360 Structured Streaming \(aka Streaming DataFrames\)](#)
- The official [Structured Streaming Programming Guide](#)
- (article) [Structured Streaming In Apache Spark](#)
- (video) [The Future of Real Time in Spark](#) from Spark Summit East 2016 in which Reynold Xin presents the concept of **Streaming DataFrames**
- (video) [Structuring Spark: DataFrames, Datasets, and Streaming](#)
- (article) [What Spark's Structured Streaming really means](#)
- (video) [A Deep Dive Into Structured Streaming](#) by Tathagata "TD" Das from Spark Summit 2016
- (video) [Arbitrary Stateful Aggregations in Structured Streaming in Apache Spark](#) by Burak Yavuz

# Batch Processing Time

**Batch Processing Time** (aka **Batch Timeout Threshold**) is the processing time (*processing timestamp*) of the current streaming batch.

The following standard functions (and their Catalyst expressions) allow accessing the batch processing time in [Micro-Batch Stream Processing](#):

- `now`, `current_timestamp`, and `unix_timestamp` functions (`CurrentTimestamp`)
- `current_date` function (`CurrentDate`)

Note

`CurrentTimestamp` or `CurrentDate` expressions are not supported in [Continuous Stream Processing](#).

## Internals

`GroupStateImpl` is given the batch processing time when [created](#) for a streaming query (that is actually the [batch processing time](#) of the `FlatMapGroupsWithStateExec` physical operator).

When created, `FlatMapGroupsWithStateExec` physical operator has the processing time undefined and set to the current timestamp in the [state preparation rule](#) every streaming batch.

The current timestamp (and other batch-specific configurations) is given as the `OffsetSeqMetadata` (as part of the query planning phase) when a [stream execution engine](#) does the following:

- `MicroBatchExecution` is requested to [construct a next streaming micro-batch](#) in [Micro-Batch Stream Processing](#)
- In [Continuous Stream Processing](#) the base `StreamExecution` is requested to [run stream processing](#) and initializes `OffsetSeqMetadata` to `0 s`.

# Internals of Streaming Queries

Note	The page is to keep notes about how to guide readers through the codebase and may disappear if merged with the other pages or become an intro page.
------	---

1. [DataStreamReader and Streaming Data Source](#)
2. [Data Source Resolution, Streaming Dataset and Logical Query Plan](#)
3. [Dataset API — High-Level DSL to Build Logical Query Plan](#)
4. [DataStreamWriter and Streaming Data Sink](#)
5. [StreamingQuery](#)
6. [StreamingQueryManager](#)

## DataStreamReader and Streaming Data Source

It all starts with `SparkSession.readStream` method which lets you define a [streaming source](#) in a **stream processing pipeline** (aka *streaming processing graph* or *dataflow graph*).

```
import org.apache.spark.sql.SparkSession
assert(spark instanceof[SparkSession])

val reader = spark.readStream

import org.apache.spark.sql.streaming.DataStreamReader
assert(reader instanceof[DataStreamReader])
```

`SparkSession.readStream` method creates a [DataStreamReader](#).

The fluent API of `DataStreamReader` allows you to describe the input data source (e.g. [DataStreamReader.format](#) and [DataStreamReader.options](#)) using method chaining (with the goal of making the readability of the source code close to that of ordinary written prose, essentially creating a domain-specific language within the interface. See [Fluent interface](#) article in Wikipedia).

```
reader
  .format("csv")
  .option("delimiter", "|")
```

There are a couple of built-in data source formats. Their names are the names of the corresponding `DataStreamReader` methods and so act like shortcuts of `DataStreamReader.format` (where you have to specify the format by name), i.e. `csv`, `json`, `orc`, `parquet` and `text`, followed by `DataStreamReader.load`.

You may also want to use `DataStreamReader.schema` method to specify the schema of the streaming data source.

```
reader.schema("a INT, b STRING")
```

In the end, you use `DataStreamReader.load` method that simply creates a streaming Dataset (the good ol' Dataset that you may have already used in Spark SQL).

```
val input = reader
  .format("csv")
  .option("delimiter", "\t")
  .schema("word STRING, num INT")
  .load("data/streaming")

import org.apache.spark.sql.DataFrame
assert(input.isInstanceOf[DataFrame])
```

The Dataset has the `isStreaming` property enabled that is basically the only way you could distinguish streaming Datasets from regular, batch Datasets.

```
assert(input.isStreaming)
```

In other words, Spark Structured Streaming is designed to extend the features of Spark SQL and let your structured queries be streaming queries.

## Data Source Resolution, Streaming Dataset and Logical Query Plan

Being curious about the internals of streaming Datasets is where you start...seeing numbers not humans (sorry, couldn't resist drawing the comparison between `Matrix the movie` and the internals of Spark Structured Streaming).

Whenever you create a Dataset (be it batch in Spark SQL or streaming in Spark Structured Streaming) is when you create a **logical query plan** using the **high-level Dataset DSL**.

A logical query plan is made up of logical operators.

Spark Structured Streaming gives you two logical operators to represent streaming sources, i.e. [StreamingRelationV2](#) and [StreamingRelation](#).

When `DataStreamReader.load` method is executed, `load` first looks up the requested data source (that you specified using `DataStreamReader.format`) and creates an instance of it (*instantiation*). That'd be **data source resolution** step (that I described in...FIXME).

`DataStreamReader.load` is where you can find the intersection of the former [Micro-Batch Stream Processing V1 API](#) with the new [Continuous Stream Processing V2 API](#).

For [MicroBatchReadSupport](#) or [ContinuousReadSupport](#) data sources,

`DataStreamReader.load` creates a logical query plan with a [StreamingRelationV2](#) leaf logical operator. That is the new **V2 code path**.

### StreamingRelationV2 Logical Operator for Data Source V2

```
// rate data source is V2
val rates = spark.readStream.format("rate").load
val plan = rates.queryExecution.logical
scala> println(plan.numberedTreeString)
00 StreamingRelationV2 org.apache.spark.sql.execution.streaming.sources.RateStreamProvider@2ed03b1a, rate, [timestamp#12, value#13L]
```

For all other types of streaming data sources, `DataStreamReader.load` creates a logical query plan with a [StreamingRelation](#) leaf logical operator. That is the former **V1 code path**.

### StreamingRelation Logical Operator for Data Source V1

```
// text data source is V1
val texts = spark.readStream.format("text").load("data/streaming")
val plan = texts.queryExecution.logical
scala> println(plan.numberedTreeString)
00 StreamingRelation DataSource(org.apache.spark.sql.SparkSession@35edd886, text, List(), None, List(), None, Map(path -> data/streaming), None), FileSource[data/streaming], [value#18]
```

## Dataset API—High-Level DSL to Build Logical Query Plan

With a streaming Dataset created, you can now use all the methods of `Dataset` API, including but not limited to the following operators:

- [Dataset.dropDuplicates](#) for streaming deduplication
- [Dataset.groupBy](#) and [Dataset.groupByKey](#) for streaming aggregation
- [Dataset.withWatermark](#) for event time watermark

Please note that a streaming Dataset is a regular Dataset (*with some streaming-related limitations*).

```
val rates = spark
  .readStream
  .format("rate")
  .load
val countByTime = rates
  .withWatermark("timestamp", "10 seconds")
  .groupBy($"timestamp")
  .agg(count("*") as "count")

import org.apache.spark.sql.Dataset
assert(countByTime.isInstanceOf[Dataset[_]])
```

The point is to understand that the Dataset API is a domain-specific language (DSL) to build a more sophisticated stream processing pipeline that you could also build using the low-level logical operators directly.

Use [Dataset.explain](#) to learn the underlying logical and physical query plans.

```

assert(countByTime.isStreaming)

scala> countByTime.explain(extended = true)
== Parsed Logical Plan ==
Aggregate ['timestamp], [unresolvedalias('timestamp, None), count(1) AS count#131L]
+- EventTimeWatermark timestamp#88: timestamp, interval 10 seconds
  +- StreamingRelationV2 org.apache.spark.sql.execution.streaming.sources.RateStreamP
rovider@2fcb3082, rate, [timestamp#88, value#89L]

== Analyzed Logical Plan ==
timestamp: timestamp, count: bigint
Aggregate [timestamp#88-T10000ms], [timestamp#88-T10000ms, count(1) AS count#131L]
+- EventTimeWatermark timestamp#88: timestamp, interval 10 seconds
  +- StreamingRelationV2 org.apache.spark.sql.execution.streaming.sources.RateStreamP
rovider@2fcb3082, rate, [timestamp#88, value#89L]

== Optimized Logical Plan ==
Aggregate [timestamp#88-T10000ms], [timestamp#88-T10000ms, count(1) AS count#131L]
+- EventTimeWatermark timestamp#88: timestamp, interval 10 seconds
  +- Project [timestamp#88]
    +- StreamingRelationV2 org.apache.spark.sql.execution.streaming.sources.RateStre
amProvider@2fcb3082, rate, [timestamp#88, value#89L]

== Physical Plan ==
*(5) HashAggregate(keys=[timestamp#88-T10000ms], functions=[count(1)], output=[timesta
mp#88-T10000ms, count#131L])
+- StateStoreSave [timestamp#88-T10000ms], state info [ checkpoint = <unknown>, runId
= 28606ba5-9c7f-4f1f-ae41-e28d75c4d948, opId = 0, ver = 0, numPartitions = 200], Append
, 0, 2
  +- *(4) HashAggregate(keys=[timestamp#88-T10000ms], functions=[merge_count(1)], out
put=[timestamp#88-T10000ms, count#136L])
    +- StateStoreRestore [timestamp#88-T10000ms], state info [ checkpoint = <unknown
>, runId = 28606ba5-9c7f-4f1f-ae41-e28d75c4d948, opId = 0, ver = 0, numPartitions = 200
], 2
      +- *(3) HashAggregate(keys=[timestamp#88-T10000ms], functions=[merge_count(1)
], output=[timestamp#88-T10000ms, count#136L])
        +- Exchange hashpartitioning(timestamp#88-T10000ms, 200)
          +- *(2) HashAggregate(keys=[timestamp#88-T10000ms], functions=[partial_
count(1)], output=[timestamp#88-T10000ms, count#136L])
            +- EventTimeWatermark timestamp#88: timestamp, interval 10 seconds
              +- *(1) Project [timestamp#88]
                +- StreamingRelation rate, [timestamp#88, value#89L]

```

Or go pro and talk to `QueryExecution` directly.

```

val plan = countByTime.queryExecution.logical
scala> println(plan.numberedTreeString)
00 'Aggregate ['timestamp], [unresolvedalias('timestamp, None), count(1) AS count#131L
]
01 +- EventTimeWatermark timestamp#88: timestamp, interval 10 seconds
02   +- StreamingRelationV2 org.apache.spark.sql.execution.streaming.sources.RateStreamProvider@2fcb3082, rate, [timestamp#88, value#89L]

```

Please note that most of the stream processing operators you may also have used in batch structured queries in Spark SQL. Again, the distinction between Spark SQL and Spark Structured Streaming is very thin from a developer's point of view.

## DataStreamWriter and Streaming Data Sink

Once you're satisfied with building a stream processing pipeline (using the APIs of [DataStreamReader](#), [Dataset](#), [RelationalGroupedDataset](#) and [KeyValueGroupedDataset](#) ), you should define how and when the result of the streaming query is persisted in (*sent out to*) an external data system using a [streaming sink](#).

Tip

Read up on the APIs of [Dataset](#), [RelationalGroupedDataset](#) and [KeyValueGroupedDataset](#) in [The Internals of Spark SQL](#) book.

You should use [Dataset.writeStream](#) method that simply creates a [DataStreamWriter](#).

```

// Not only is this a Dataset, but it is also streaming
assert(countByTime.isStreaming)

val writer = countByTime.writeStream

import org.apache.spark.sql.streaming.DataStreamWriter
assert(writer.asInstanceOf[DataStreamWriter[_]])

```

The fluent API of [DataStreamWriter](#) allows you to describe the output data sink (e.g. [DataStreamWriter.format](#) and [DataStreamWriter.options](#)) using method chaining (with the goal of making the readability of the source code close to that of ordinary written prose, essentially creating a domain-specific language within the interface. See [Fluent interface](#) article in Wikipedia).

```

writer
  .format("csv")
  .option("delimiter", "\t")

```

Like in [DataStreamReader](#) data source formats, there are a couple of built-in data sink formats. Unlike data source formats, their names do not have corresponding `DataStreamWriter` methods. The reason is that you will use [DataStreamWriter.start](#) to create and immediately start a [StreamingQuery](#).

There are however two special output formats that do have corresponding `DataStreamWriter` methods, i.e. [DataStreamWriter.foreach](#) and [DataStreamWriter.foreachBatch](#), that allow for persisting query results to external data systems that do not have streaming sinks available. They give you a trade-off between developing a full-blown streaming sink and simply using the methods (that lay the basis of what a custom sink would have to do anyway).

`DataStreamWriter` API defines two new concepts (that are not available in the "base" Spark SQL):

- [OutputMode](#) that you specify using [DataStreamWriter.outputMode](#) method
- [Trigger](#) that you specify using [DataStreamWriter.trigger](#) method

You may also want to give a streaming query a name using [DataStreamWriter.queryName](#) method.

In the end, you use [DataStreamWriter.start](#) method to create and immediately start a [StreamingQuery](#).

```
import org.apache.spark.sql.streaming.OutputMode
import org.apache.spark.sql.streaming.Trigger
import scala.concurrent.duration._
val sq = writer
  .format("console")
  .option("truncate", false)
  .option("checkpointLocation", "/tmp/csv-to-csv-checkpoint")
  .outputMode(OutputMode.Append)
  .trigger(Trigger.ProcessingTime(30.seconds))
  .queryName("csv-to-csv")
  .start("/tmp")

import org.apache.spark.sql.streaming.StreamingQuery
assert(sq.isInstanceOf[StreamingQuery])
```

When `DataStreamWriter` is requested to [start a streaming query](#), it allows for the following data source formats:

- **memory** with [MemorySinkV2](#) (with [ContinuousTrigger](#)) or [MemorySink](#)
- **foreach** with [ForeachWriterProvider](#) sink

- **foreachBatch** with [ForeachBatchSink](#) sink (that does not support [ContinuousTrigger](#))
- Any `DataSourceRegister` data source
- Custom data sources specified by their fully-qualified class names or `[name].DefaultSource`
- **avro, kafka and some others** (see `DataSource.lookupDataSource` object method)
- [StreamWriter](#)
- `DataSource` is requested to [create a streaming sink](#) that accepts [StreamSinkProvider](#) or `FileFormat` data sources only

With a streaming sink, `DataStreamWriter` requests the [StreamingQueryManager](#) to [start a streaming query](#).

## StreamingQuery

When a stream processing pipeline is started (using `DataStreamWriter.start` method), `DataStreamWriter` creates a [StreamingQuery](#) and requests the [StreamingQueryManager](#) to [start a streaming query](#).

## StreamingQueryManager

[StreamingQueryManager](#) is used to manage streaming queries.

# Streaming Join

In Spark Structured Streaming, a **streaming join** is a streaming query that was described (*build*) using the [high-level streaming operators](#):

- [Dataset.crossJoin](#)
- [Dataset.join](#)
- [Dataset.joinWith](#)
- SQL's `JOIN` clause

Streaming joins can be **stateless** or **stateful**:

- Joins of a streaming query and a batch query (*stream-static joins*) are stateless and no state management is required
- Joins of two streaming queries ([stream-stream joins](#)) are stateful and require streaming state (with an optional [join state watermark for state removal](#)).

## Stream-Stream Joins

Spark Structured Streaming supports **stream-stream joins** with the following:

- [Equality predicate](#) (i.e. [equi-joins](#) that use only equality comparisons in the join predicate)
- `Inner` , `LeftOuter` , and `RightOuter` [join types only](#)

Stream-stream equi-joins are planned as [StreamingSymmetricHashJoinExec](#) physical operators of two [ShuffleExchangeExec](#) physical operators (per [Required Partition Requirements](#)).

## Join State Watermark for State Removal

Stream-stream joins may optionally define **Join State Watermark** for state removal (cf. [Watermark Predicates for State Removal](#)).

A join state watermark can be specified on the following:

1. [Join keys \(key state\)](#)
2. [Columns of the left and right sides \(value state\)](#)

A join state watermark can be specified on key state, value state or both.

## IncrementalExecution — QueryExecution of Streaming Queries

Under the covers, the [high-level operators](#) create a logical query plan with one or more `Join` logical operators.

**Tip** [Read up on Join Logical Operator](#) in [The Internals of Spark SQL](#) online book.

In Spark Structured Streaming [IncrementalExecution](#) is responsible for planning streaming queries for execution.

At [query planning](#), `IncrementalExecution` uses the [StreamingJoinStrategy](#) execution planning strategy for planning [stream-stream joins](#) as [StreamingSymmetricHashJoinExec](#) physical operators.

## Demos

Use the following demo application to learn more:

- [StreamStreamJoinApp](#)

## Further Reading Or Watching

- [Stream-stream Joins](#) in the official documentation of Apache Spark for Structured Streaming
- [Introducing Stream-Stream Joins in Apache Spark 2.3](#) by Databricks
- (video) [Deep Dive into Stateful Stream Processing in Structured Streaming](#) by Tathagata Das

# StateStoreAwareZipPartitionsRDD

`StateStoreAwareZipPartitionsRDD` is a `ZippedPartitionsRDD2` with the `left` and `right` parent RDDs.

`StateStoreAwareZipPartitionsRDD` is created exclusively when `StreamingSymmetricHashJoinExec` physical operator is requested to execute and generate a recipe for a distributed computation (as an `RDD[InternalRow]`) (and requests `StateStoreAwareZipPartitionsHelper` for one).

## Creating StateStoreAwareZipPartitionsRDD Instance

`StateStoreAwareZipPartitionsRDD` takes the following to be created:

- `SparkContext`
- Function (`(Iterator[A], Iterator[B]) => Iterator[V]`, e.g. `processPartitions`)
- **Left RDD** - the RDD of the left side of a join (`RDD[A]`)
- **Right RDD** - the RDD of the right side of a join (`RDD[B]`)
- `StatefulOperatorStateInfo`
- Names of the `state stores`
- `StateStoreCoordinatorRef`

## Placement Preferences of Partition (Preferred Locations) — `getPreferredLocations` Method

```
getPreferredLocations(partition: Partition): Seq[String]
```

Note	<code>getPreferredLocations</code> is a part of the RDD Contract to specify placement preferences (aka <i>preferred task locations</i> ), i.e. where tasks should be executed to be as close to the data as possible.
------	---

`getPreferredLocations` simply requests the `StateStoreCoordinatorRef` for the location of every `state store` (with the `StatefulOperatorStateInfo` and the partition ID) and returns unique executor IDs (so that processing a partition happens on the executor with the proper state store for the operator and the partition).



# SymmetricHashJoinStateManager

`SymmetricHashJoinStateManager` is created for the left and right `OneSideHashJoiners` of a `StreamingSymmetricHashJoinExec` physical operator (one for each side when `StreamingSymmetricHashJoinExec` is requested to process partitions of the left and right sides of a stream-stream join).

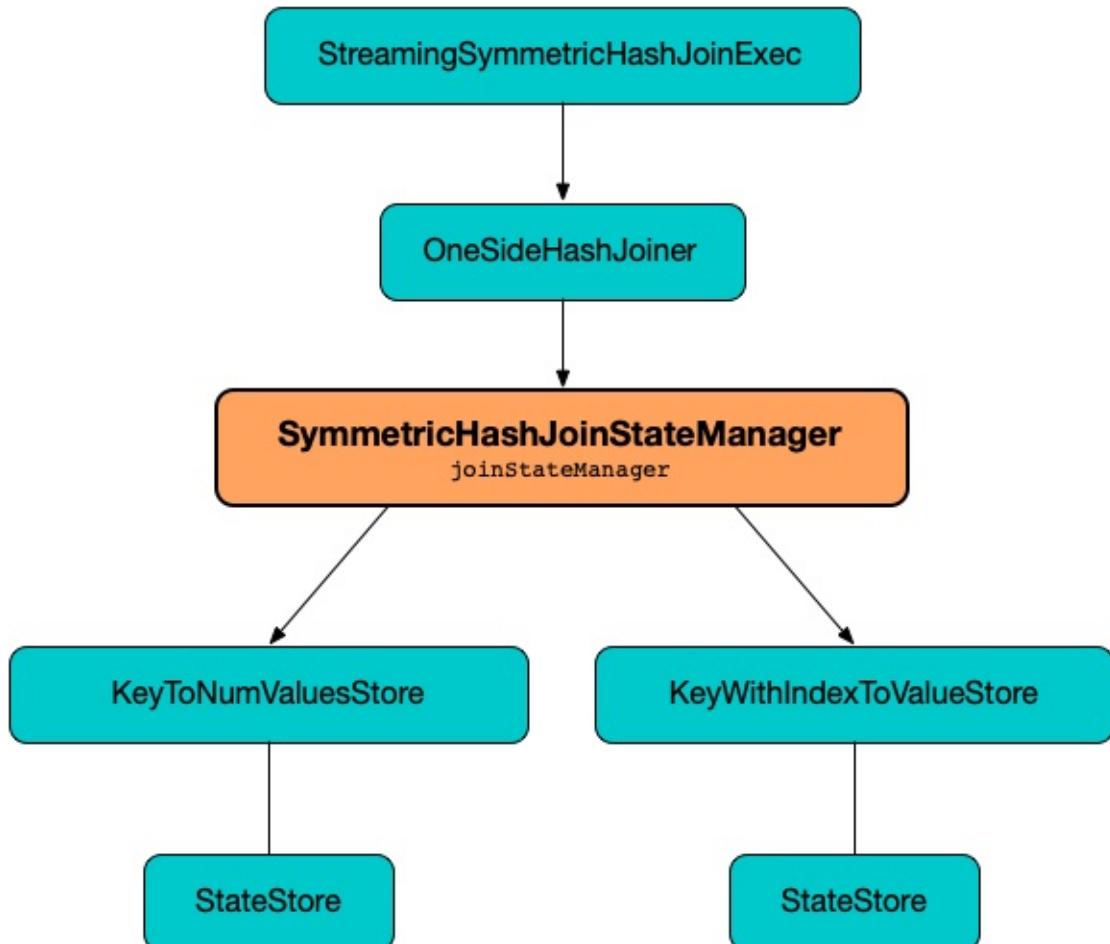


Figure 1. SymmetricHashJoinStateManager and Stream-Stream Join

`SymmetricHashJoinStateManager` manages join state using the `KeyToNumValuesStore` and the `KeyWithIndexToValueStore` state store handlers (and simply acts like their facade).

## Creating SymmetricHashJoinStateManager Instance

`SymmetricHashJoinStateManager` takes the following to be created:

- `JoinSide`
- Attributes of input values
- Join keys ( `Seq[Expression]` )

- `StatefulOperatorStateInfo`
- `StateStoreConf`
- Hadoop Configuration

`SymmetricHashJoinStateManager` initializes the internal properties.

## KeyToNumValuesStore and KeyWithIndexToValueStore State Store Handlers — `keyToNumValues` and `keyWithIndexToValue` Internal Properties

`SymmetricHashJoinStateManager` uses a `KeyToNumValuesStore` (`keyToNumValues`) and a `KeyWithIndexToValueStore` (`keyWithIndexToValue`) internally that are created immediately when `SymmetricHashJoinStateManager` is created (for a `OneSideHashJoiner`).

`keyToNumValues` and `keyWithIndexToValue` are used when `SymmetricHashJoinStateManager` is requested for the following:

- Retrieving the value rows by key
- Append a new value row to a given key
- `removeByKeyCondition`
- `removeByValueCondition`
- Commit state changes
- Abort state changes
- Performance metrics

## Join Side Marker — `JoinSide` Internal Enum

`JoinSide` can be one of the two possible values:

- `LeftSide` (alias: `left`)
- `RightSide` (alias: `right`)

They are both used exclusively when `StreamingSymmetricHashJoinExec` binary physical operator is requested to `execute` (and process partitions of the left and right sides of a stream-stream join with an `OneSideHashJoiner`).

## Performance Metrics — `metrics` Method

```
metrics: StateStoreMetrics
```

`metrics` returns the combined [StateStoreMetrics](#) of the [KeyToNumValuesStore](#) and the [KeyWithIndexToValueStore](#) state store handlers.

**Note**

`metrics` is used exclusively when `OneSideHashJoiner` is requested to [commitStateAndGetMetrics](#).

## removeByKeyCondition Method

```
removeByKeyCondition(  
    removalCondition: UnsafeRow => Boolean): Iterator[UnsafeRowPair]
```

`removeByKeyCondition` creates an `Iterator` of `UnsafeRowPairs` that removes keys (and associated values) for which the given `removalCondition` predicate holds.

`removeByKeyCondition` uses the [KeyToNumValuesStore](#) for all state keys and values (in the underlying state store).

**Note**

`removeByKeyCondition` is used exclusively when `OneSideHashJoiner` is requested to [remove an old state](#) (for [JoinStateKeyWatermarkPredicate](#)).

## getNext Internal Method (of removeByKeyCondition Method)

```
getNext(): UnsafeRowPair
```

`getNext` goes over the keys and values in the [allKeyToNumValues](#) sequence and removes keys (from the [KeyToNumValuesStore](#)) and the corresponding values (from the [KeyWithIndexToValueStore](#)) for which the given `removalCondition` predicate holds.

## removeByValueCondition Method

```
removeByValueCondition(  
    removalCondition: UnsafeRow => Boolean): Iterator[UnsafeRowPair]
```

`removeByValueCondition` creates an `Iterator` of `UnsafeRowPairs` that removes values (and associated keys if needed) for which the given `removalCondition` predicate holds.

Note	<code>removeByValueCondition</code> is used exclusively when <code>OneSideHashJoiner</code> is requested to <a href="#">remove an old state</a> (when <code>JoinStateValueWatermarkPredicate</code> is used).
------	---

## getNext Internal Method (of removeByValueCondition Method)

```
getNext(): UnsafeRowPair
```

```
getNext ...FIXME
```

## Appending New Value Row to Key — append Method

```
append(  
    key: UnsafeRow,  
    value: UnsafeRow): Unit
```

`append` requests the [KeyToNumValuesStore](#) for the number of value rows for the given key.

In the end, `append` requests the stores for the following:

- [KeyWithIndexToValueStore](#) to store the given value row
- [KeyToNumValuesStore](#) to store the given key with the number of value rows incremented.

Note	<code>append</code> is used exclusively when <code>OneSideHashJoiner</code> is requested to <a href="#">storeAndJoinWithOtherSide</a> .
------	---

## Retrieving Value Rows By Key — get Method

```
get(key: UnsafeRow): Iterator[UnsafeRow]
```

`get` requests the [KeyToNumValuesStore](#) for the number of value rows for the given key.

In the end, `get` requests the [KeyWithIndexToValueStore](#) to retrieve that number of value rows for the given key and leaves value rows only.

Note	<code>get</code> is used when <code>OneSideHashJoiner</code> is requested to <a href="#">storeAndJoinWithOtherSide</a> and <a href="#">retrieving value rows for a key</a> .
------	--

## Committing State (Changes) — `commit` Method

```
commit(): Unit
```

`commit` simply requests the `keyToNumValues` and `keyWithIndexToValue` state store handlers to `commit state changes`.

Note

`commit` is used exclusively when `OneSideHashJoiner` is requested to `commit state changes and get performance metrics`.

## Aborting State (Changes) — `abortIfNeeded` Method

```
abortIfNeeded(): Unit
```

`abortIfNeeded` ...FIXME

Note

`abortIfNeeded` is used when...FIXME

## allStateStoreNames Object Method

```
allStateStoreNames(joinSides: JoinSide*): Seq[String]
```

`allStateStoreNames` simply returns the `names of the state stores` for all possible combinations of the given `JoinSides` and the two possible store types (e.g. `keyToNumValues` and `keyWithIndexToValue`).

Note

`allStateStoreNames` is used exclusively when `StreamingSymmetricHashJoinExec` physical operator is requested to `execute and generate the runtime representation` (as a `RDD[InternalRow]` ).

## getStateStoreName Object Method

```
getStateStoreName(  
  joinSide: JoinSide,  
  storeType: StateStoreType): String
```

`getStateStoreName` simply returns a string of the following format:

```
[joinSide]-[storeType]
```

Note	<p><code>getStateStoreName</code> is used when:</p> <ul style="list-style-type: none"> <li>• <code>StateStoreHandler</code> is requested to <a href="#">load a state store</a></li> <li>• <code>SymmetricHashJoinStateManager</code> utility is requested for <a href="#">allStateStoreNames</a> (for <code>StreamingSymmetricHashJoinExec</code> physical operator to <a href="#">execute</a> and <a href="#">generate the runtime representation</a>)</li> </ul>
------	--

## updateNumValueForCurrentKey Internal Method

`updateNumValueForCurrentKey(): Unit`

`updateNumValueForCurrentKey` ...FIXME

Note	<p><code>updateNumValueForCurrentKey</code> is used exclusively when <code>SymmetricHashJoinStateManager</code> is requested to <a href="#">removeByValueCondition</a>.</p>
------	---

## Internal Properties

Name	Description
<code>keyAttributes</code>	<p>Key attributes, i.e. <code>AttributeReferences</code> of the <a href="#">key schema</a>          Used exclusively in <code>KeyWithIndexToValueStore</code> when requested for the <code>keyWithIndexExprs</code>, <code>indexOrdinalInKeyWithIndexRow</code>, <code>keyWithIndexRowGenerator</code> and <code>keyRowGenerator</code></p>
<code>keySchema</code>	<p>Key schema (<code>structType</code>) based on the <a href="#">join keys</a> with the names in the format of <b>field</b> and their ordinals (index)          Used when:</p> <ul style="list-style-type: none"> <li>• <code>SymmetricHashJoinStateManager</code> is requested for the <a href="#">key attributes</a> (for <code>KeyWithIndexToValueStore</code>)</li> <li>• <code>KeyToNumValuesStore</code> is requested for the <a href="#">state store</a></li> <li>• <code>KeyWithIndexToValueStore</code> is requested for the <a href="#">keyWithIndexSchema</a> (for the internal <a href="#">state store</a>)</li> </ul>

# StateStoreHandler Internal Contract

`StateStoreHandler` is the internal base of state store handlers that manage a `StateStore` (i.e. `commit`, `abortIfNeeded` and `metrics`).

`StateStoreHandler` takes a single `stateStoreType` to be created:

- `KeyToNumValuesType` for `KeyToNumValuesStore` (alias: `keyToNumValues`)
- `KeyWithIndexToValueType` for `KeyWithIndexToValueStore` (alias: `keyWithIndexToValue`)

**Note**

`StateStoreHandler` is a Scala private abstract class and cannot be created directly. It is created indirectly for the concrete StateStoreHandlers.

Table 1. StateStoreHandler Contract

Method	Description
<code>stateStore</code>	<code>stateStore: StateStore</code>  <code>StateStore</code>

Table 2. StateStoreHandlers

StateStoreHandler	Description
<code>KeyToNumValuesStore</code>	<code>StateStoreHandler</code> of <code>KeyToNumValuesType</code>
<code>KeyWithIndexToValueStore</code>	

**Tip**

Enable `ALL` logging levels for  
`org.apache.spark.sql.execution.streaming.state.SymmetricHashJoinStateManager.State`  
happens inside.

Add the following line to `conf/log4j.properties` :

```
log4j.logger.org.apache.spark.sql.execution.streaming.state.SymmetricHashJoinSta
```

Refer to [Logging](#).

## Performance Metrics — metrics Method

`metrics: StateStoreMetrics`

`metrics` simply requests the [StateStore](#) for the [StateStoreMetrics](#).

**Note**

`metrics` is used exclusively when [SymmetricHashJoinStateManager](#) is requested for the [metrics](#).

## Committing State (Changes to State Store) — `commit` Method

```
commit(): Unit
```

`commit` ...FIXME

**Note**

`commit` is used when...FIXME

## `abortIfNeeded` Method

```
abortIfNeeded(): Unit
```

`abortIfNeeded` ...FIXME

**Note**

`abortIfNeeded` is used when...FIXME

## Loading State Store (By Key and Value Schemas) — `getStateStore` Method

```
getStateStore(  
    keySchema: StructType,  
    valueSchema: StructType): StateStore
```

`getStateStore` creates a new [StateStoreProviderId](#) (for the [StatefulOperatorStateInfo](#) of the owning [SymmetricHashJoinStateManager](#), the partition ID from the execution context, and the [name of the state store](#) for the [JoinSide](#) and [StateStoreType](#)).

`getStateStore` uses the `stateStore` utility to look up a [StateStore](#) for the [StateStoreProviderId](#).

In the end, `getStateStore` prints out the following INFO message to the logs:

```
Loaded store [storeId]
```

**Note**

`getStateStore` is used when [KeyToNumValuesStore](#) and [KeyWithIndexToValueStore](#) state store handlers are created (for [SymmetricHashJoinStateManager](#)).

## StateStoreType Contract (Sealed Trait)

`StateStoreType` is required to create a [StateStoreHandler](#).

Table 3. StateStoreTypes

StateStoreType	toString	Description
<code>KeyToNumValuesType</code>	<code>keyToNumValues</code>	
<code>KeyWithIndexToValueType</code>	<code>keyWithIndexToValue</code>	

**Note**

`StateStoreType` is a Scala private **sealed trait** which means that all the [implementations](#) are in the same compilation unit (a single file).

# KeyToNumValuesStore — State Store (Handler) Of Join Keys And Counts

`KeyToNumValuesStore` is a `StateStoreHandler` (of `KeyToNumValueType`) for `SymmetricHashJoinStateManager` to manage a `join state`.

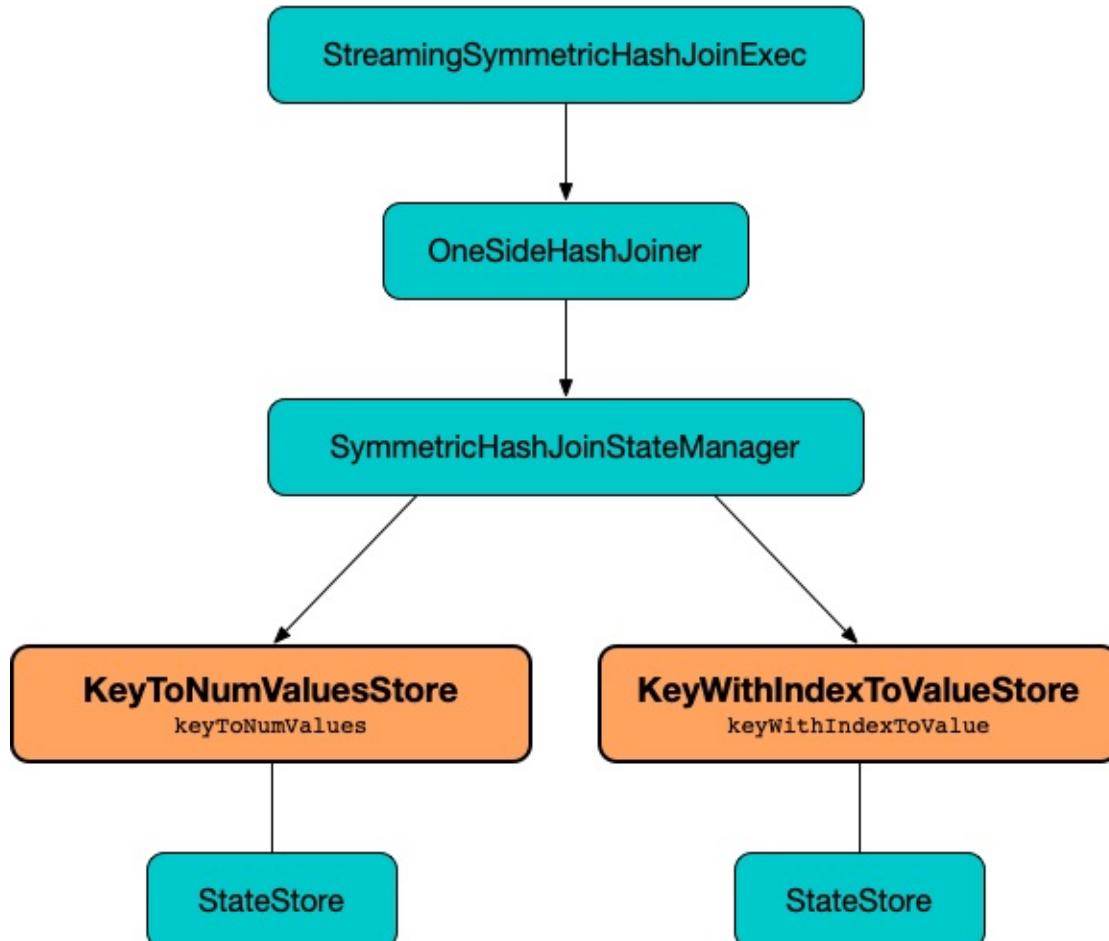


Figure 1. `KeyToNumValuesStore`, `KeyWithIndexToValueStore` and Stream-Stream Join As a `StateStoreHandler`, `KeyToNumValuesStore` manages a `state store` (that is `loaded`) with the join keys (per `key schema`) and their count (per `value schema`).

`KeyToNumValuesStore` uses the schema for values in the `state store` with one field `value` (of type `long`) that is the number of value rows (count).

Enable ALL logging level for `org.apache.spark.sql.execution.streaming.state.SymmetricHashJoinStateManager$KeyToNumValuesStore` to see what happens inside.

Add the following line to `conf/log4j.properties`:

Tip

```
log4j.logger.org.apache.spark.sql.execution.streaming.state.SymmetricHashJoinStateManager$KeyToNumValuesStore=ALL
```

Refer to [Logging](#).

## Looking Up Number Of Value Rows For Given Key (Value Count) — `get` Method

```
get(key: UnsafeRow): Long
```

`get` requests the [StateStore](#) for the value for the given key and returns the long value at 0 th position (of the row found) or 0 .

Note

`get` is used when `SymmetricHashJoinStateManager` is requested for the values for a given key and append a new value to a given key.

## Storing Key Count For Given Key — `put` Method

```
put(
  key: UnsafeRow,
  numValues: Long): Unit
```

`put` stores the `numValues` at the 0 th position (of the internal unsafe row) and requests the [StateStore](#) to store it with the given key.

`put` requires that the `numValues` count is greater than 0 (or throws an `IllegalArgumentException` ).

Note

`put` is used when `SymmetricHashJoinStateManager` is requested for the append a new value to a given key and `updateNumValueForCurrentKey`.

## All State Keys and Values — `iterator` Method

```
iterator: Iterator[KeyAndNumValues]
```

`iterator` simply requests the [StateStore](#) for all state keys and values.

Note	<code>iterator</code> is used when <code>SymmetricHashJoinStateManager</code> is requested to <a href="#">removeByKeyCondition</a> and <a href="#">removeByValueCondition</a> .
------	---

## Removing State Key — `remove` Method

```
remove(key: UnsafeRow): Unit
```

`remove` simply requests the [StateStore](#) to [remove](#) the given key.

Note	<code>remove</code> is used when...FIXME
------	--

# KeyWithIndexToValueStore — State Store (Handler) Of Join Keys With Index Of Values

`KeyWithIndexToValueStore` is a [StateStoreHandler](#) (of `KeyWithIndexValueType`) for [SymmetricHashJoinStateManager](#) to manage a [join state](#).

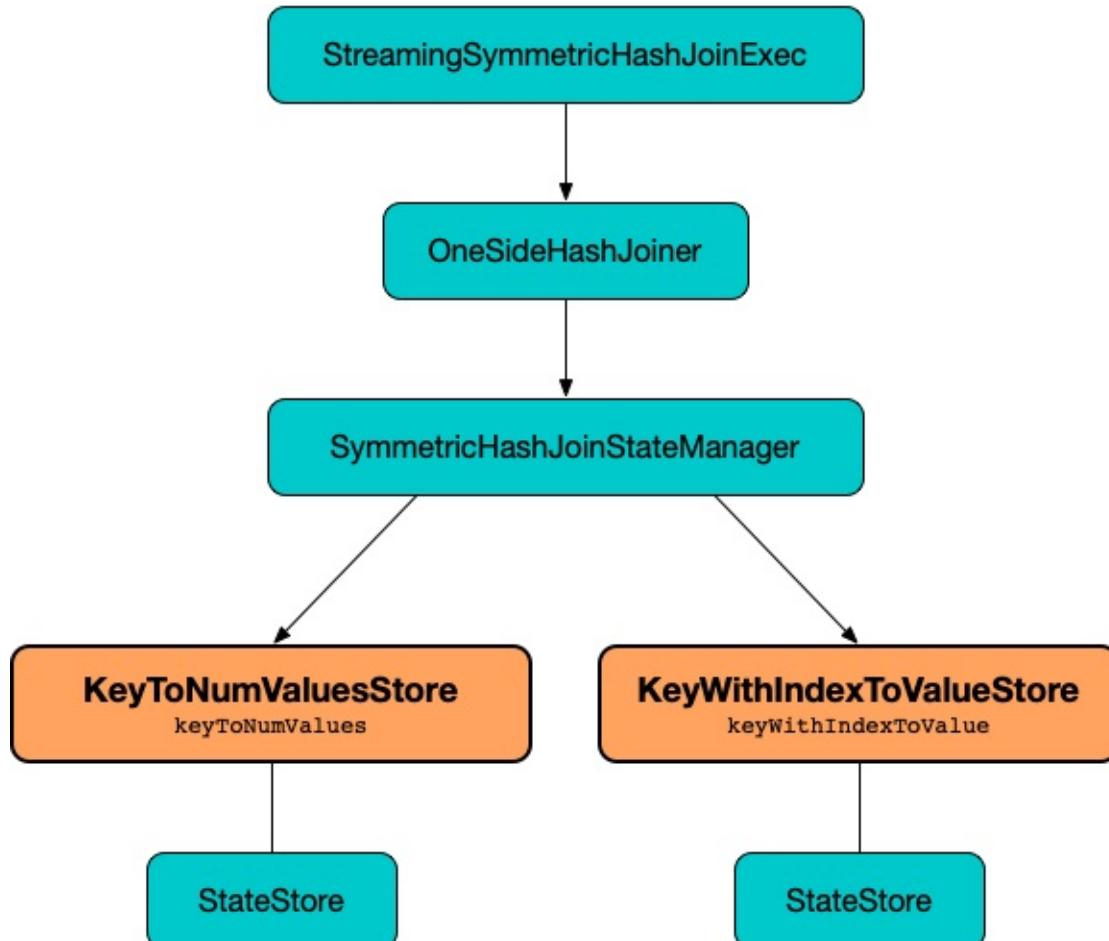


Figure 1. `KeyToNumValuesStore`, `KeyWithIndexToValueStore` and Stream-Stream Join As a [StateStoreHandler](#), `KeyWithIndexToValueStore` manages a [state store](#) (that is [loaded](#)) for keys and values per the [keys with index](#) and [input values](#) schemas, respectively.

`KeyWithIndexToValueStore` uses a schema (for the [state store](#)) that is the [key schema](#) (of the parent `SymmetricHashJoinStateManager`) with an extra field `index` of type `long`.

Enable ALL logging level for `org.apache.spark.sql.execution.streaming.state.SymmetricHashJoinStateManager$KeyWithIndexToValueStore` to see what happens inside.

Add the following line to `conf/log4j.properties`:

Tip

```
log4j.logger.org.apache.spark.sql.execution.streaming.state.SymmetricHashJoinSta
```

Refer to [Logging](#).

## Looking Up State Row For Given Key and Index — `get` Method

```
get(
  key: UnsafeRow,
  valueIndex: Long): UnsafeRow
```

`get` simply requests the internal `state store` to look up the value for the given `key` and `valueIndex`.

Note

`get` is used exclusively when `SymmetricHashJoinStateManager` is requested to `removeByValueCondition`

## Retrieving (Given Number of) Values for Key — `getAll` Method

```
getAll(
  key: UnsafeRow,
  numValues: Long): Iterator[KeyWithIndexAndValue]
```

`getAll` ...FIXME

Note

`getAll` is used when `SymmetricHashJoinStateManager` is requested to get values for a given key and `removeByKeyCondition`.

## Storing State Row For Given Key and Index — `put` Method

```
put(
  key: UnsafeRow,
  valueIndex: Long,
  value: UnsafeRow): Unit
```

put ...FIXME

Note

`put` is used when `SymmetricHashJoinStateManager` is requested to append a new value to a given key and `removeByKeyCondition`.

## remove Method

```
remove(
  key: UnsafeRow,
  valueIndex: Long): Unit
```

remove ...FIXME

Note

`remove` is used when `SymmetricHashJoinStateManager` is requested to `removeByKeyCondition` and `removeByValueCondition`.

## keyWithIndexRow Internal Method

```
keyWithIndexRow(
  key: UnsafeRow,
  valueIndex: Long): UnsafeRow
```

`keyWithIndexRow` uses the `keyWithIndexRowGenerator` to generate an `UnsafeRow` for the `key` and sets the `valueIndex` at the `indexOrdinalInKeyWithIndexRow` position.

Note

`keyWithIndexRow` is used when `KeyWithIndexToValueStore` is requested to `get`, `getAll`, `put`, `remove` and `removeAllValues`.

## removeAllValues Method

```
removeAllValues(
  key: UnsafeRow,
  numValues: Long): Unit
```

`removeAllValues` ...FIXME

Note	<code>removeAllValues</code> does not seem to be used at all.
------	---

## iterator Method

```
iterator: Iterator[KeyWithIndexAndValue]
```

```
iterator ...FIXME
```

Note	<code>iterator</code> does not seem to be used at all.
------	--

## Internal Properties

Name	Description
<code>indexOrdinalInKeyWithIndexRow</code>	Position of the index in the key row (which corresponds to the number of the <a href="#">key attributes</a> ) Used exclusively in the <a href="#">keyWithIndexRow</a>
<code>keyWithIndexExprs</code>	<a href="#">keyAttributes</a> with <code>Literal(1L)</code> expression appended Used exclusively for the <a href="#">keyWithIndexRowGenerator</a> projection
<code>keyWithIndexRowGenerator</code>	<code>UnsafeProjection</code> for the <a href="#">keyWithIndexExprs</a> bound to the <a href="#">keyAttributes</a> Used exclusively in <a href="#">keyWithIndexRow</a>

# OneSideHashJoiner

`OneSideHashJoiner` manages join state of one side of a [stream-stream join](#) (using [SymmetricHashJoinStateManager](#)).

`OneSideHashJoiner` is [created](#) exclusively for `StreamingSymmetricHashJoinExec` physical operator (when requested to [process partitions](#) of the left and right sides of a stream-stream join).

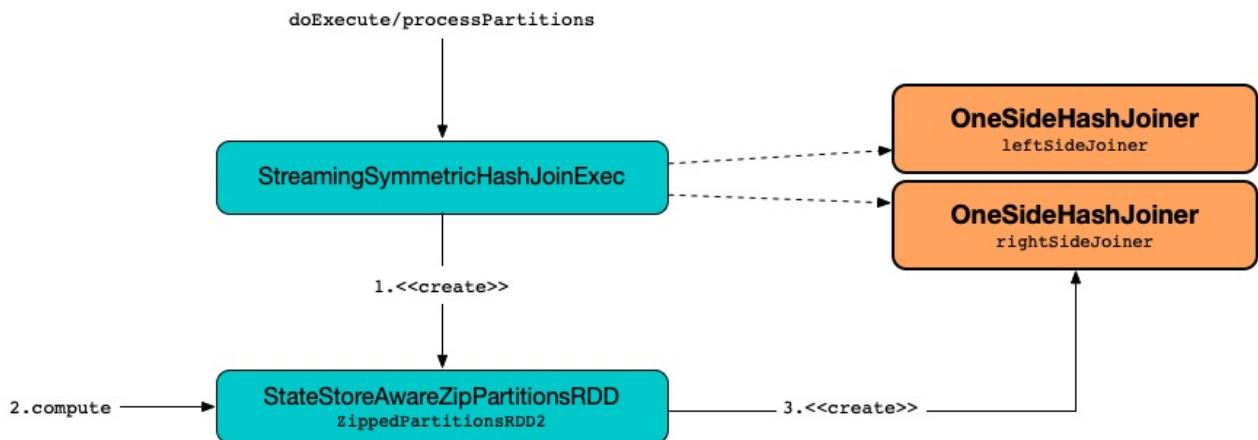


Figure 1. OneSideHashJoiner and StreamingSymmetricHashJoinExec

`StreamingSymmetricHashJoinExec` physical operator uses two `OneSideHashJoiner`s per side of the stream-stream join ([left](#) and [right](#) sides).

`OneSideHashJoiner` uses an [optional join state watermark predicate](#) to remove old state.

## Note

`OneSideHashJoiner` is a Scala private internal class of `StreamingSymmetricHashJoinExec` and so has full access to `StreamingSymmetricHashJoinExec` properties.

## Creating OneSideHashJoiner Instance

`OneSideHashJoiner` takes the following to be created:

- `JoinSide`
- Input attributes ( `Seq[Attribute]` )
- Join keys ( `Seq[Expression]` )
- Input rows ( `Iterator[InternalRow]` )
- Optional pre-join filter Catalyst expression
- Post-join filter ( `(InternalRow) => Boolean` )

- `JoinStateWatermarkPredicate`

`OneSideHashJoiner` initializes the internal registries and counters.

## SymmetricHashJoinStateManager — `joinStateManager` Internal Property

```
joinStateManager: SymmetricHashJoinStateManager
```

`joinStateManager` is a `SymmetricHashJoinStateManager` that is created for a `OneSideHashJoiner` (with the join side, the `input attributes`, the `join keys`, and the `StatefulOperatorStateInfo` of the owning `StreamingSymmetricHashJoinExec`).

`joinStateManager` is used when `OneSideHashJoiner` is requested for the following:

- `storeAndJoinWithOtherSide`
- Get the values for a given key
- Remove an old state
- `commitStateAndGetMetrics`

## Number of Updated State Rows — `updatedStateRowsCount` Internal Counter

`updatedStateRowsCount` is the number the join keys and associated rows that were persisted as a join state, i.e. how many times `storeAndJoinWithOtherSide` requested the `SymmetricHashJoinStateManager` to append the join key and the input row (to a join state).

`updatedStateRowsCount` is then used (via `numUpdatedStateRows` method) for the `numUpdatedStateRows` performance metric.

`updatedStateRowsCount` is available via `numUpdatedStateRows` method.

```
numUpdatedStateRows: Long
```

Note

`numUpdatedStateRows` is used exclusively when `StreamingSymmetricHashJoinExec` physical operator is requested to [spark-sql-streaming-`StreamingSymmetricHashJoinExec#processPartitions` process partitions of the left and right sides of a stream-stream join] (and completes).

## Optional Join State Watermark Predicate — stateWatermarkPredicate Internal Property

```
stateWatermarkPredicate: Option[JoinStateWatermarkPredicate]
```

When `created`, `OneSideHashJoiner` is given a `JoinStateWatermarkPredicate`.

`stateWatermarkPredicate` is used for the `stateKeyWatermarkPredicateFunc` (when a `JoinStateKeyWatermarkPredicate`) and the `stateValueWatermarkPredicateFunc` (when a `JoinStateValueWatermarkPredicate`) that are both used when `OneSideHashJoiner` is requested to `removeOldState`.

## storeAndJoinWithOtherSide Method

```
storeAndJoinWithOtherSide(  
    otherSideJoiner: OneSideHashJoiner)(  
    generateJoinedRow: (InternalRow, InternalRow) => JoinedRow): Iterator[InternalRow]
```

`storeAndJoinWithOtherSide` tries to find the `watermark attribute` among the `input attributes`.

`storeAndJoinWithOtherSide` creates a `watermark expression` (for the `watermark attribute` and the current `event-time watermark`).

With the `watermark attribute` found, `storeAndJoinWithOtherSide` generates a new predicate for the `watermark expression` and the `input attributes` that is then used to filter out (`exclude`) late rows from the `input`. Otherwise, the input rows are left unchanged (i.e. no rows are considered late and excluded).

For every `input row` (possibly `watermarked`), `storeAndJoinWithOtherSide` applies the `preJoinFilter` predicate and branches off per result (`true` or `false`).

Note

`storeAndJoinWithOtherSide` is used when `StreamingSymmetricHashJoinExec` physical operator is requested to process partitions of the left and right sides of a stream-stream join.

## preJoinFilter Predicate Positive ( true )

When the `preJoinFilter` predicate succeeds on an input row, `storeAndJoinWithOtherSide` extracts the join key (using the `keyGenerator`) and requests the given `OneSideHashJoiner` (`otherSideJoiner`) for the `SymmetricHashJoinStateManager` that is in turn requested for the state values for the extracted join key. The values are then processed (*mapped over*) using the given `generateJoinedRow` function and then filtered by the `post-join filter`.

`storeAndJoinWithOtherSide` uses the `stateKeyWatermarkPredicateFunc` (on the extracted join key) and the `stateValueWatermarkPredicateFunc` (on the current input row) to determine whether to request the `SymmetricHashJoinStateManager` to `append` the key and the input row (to a join state). If so, `storeAndJoinWithOtherSide` increments the `updatedStateRowsCount` counter.

## preJoinFilter Predicate Negative ( false )

When the `preJoinFilter` predicate fails on an input row, `storeAndJoinWithOtherSide` creates a new `Iterator[InternalRow]` of joined rows per `join side` and `type`:

- For `LeftSide` and `Leftouter`, the join row is the current row with the values of the right side all `null` (`nullRight`)
- For `RightSide` and `Rightouter`, the join row is the current row with the values of the left side all `null` (`nullLeft`)
- For all other combinations, the iterator is simply empty (that will be removed from the output by the outer `nonLateRows.flatMap`).

## Removing Old State— removeOldState Method

```
removeOldState(): Iterator[UnsafeRowPair]
```

`removeOldState` branches off per the `JoinStateWatermarkPredicate`:

- For `JoinStateKeyWatermarkPredicate`, `removeOldState` requests the `SymmetricHashJoinStateManager` to `removeByKeyCondition` (with the `stateKeyWatermarkPredicateFunc`)
- For `JoinStateValueWatermarkPredicate`, `removeOldState` requests the `SymmetricHashJoinStateManager` to `removeByValueCondition` (with the `stateValueWatermarkPredicateFunc`)
- For any other predicates, `removeOldState` returns an empty iterator (no rows to process)

Note

`removeOldState` is used exclusively when `StreamingSymmetricHashJoinExec` physical operator is requested to process partitions of the left and right sides of a stream-stream join.

## Retrieving Value Rows For Key— get Method

```
get(key: UnsafeRow): Iterator[UnsafeRow]
```

`get` simply requests the [SymmetricHashJoinStateManager](#) to retrieve value rows for the [key](#).

Note

`get` is used exclusively when `StreamingSymmetricHashJoinExec` physical operator is requested to process partitions of the left and right sides of a stream-stream join.

## Committing State (Changes) and Requesting Performance Metrics — `commitStateAndGetMetrics` Method

```
commitStateAndGetMetrics(): StateStoreMetrics
```

`commitStateAndGetMetrics` simply requests the [SymmetricHashJoinStateManager](#) to commit followed by requesting for the [performance metrics](#).

Note

`commitStateAndGetMetrics` is used exclusively when `StreamingSymmetricHashJoinExec` physical operator is requested to process partitions of the left and right sides of a stream-stream join.

## Internal Properties

Name	Description
keyGenerator	<p>keyGenerator: UnsafeProjection</p> <p>Function to project (<i>extract</i>) join keys from an input row.</p> <p>Used when...FIXME</p>
preJoinFilter	<p>preJoinFilter: InternalRow =&gt; Boolean</p> <p>Used when...FIXME</p>
stateKeyWatermarkPredicateFunc	<p>stateKeyWatermarkPredicateFunc: InternalRow =&gt; Boolean</p> <p>Predicate for late rows based on the <a href="#">stateWatermarkPredicate</a></p> <p>Used for the following:</p> <ul style="list-style-type: none"> <li>• <a href="#">storeAndJoinWithOtherSide</a> (and check out where to append a row to the <a href="#">SymmetricHashJoinStateM</a>)</li> <li>• <a href="#">removeOldState</a></li> </ul>
stateValueWatermarkPredicateFunc	<p>stateValueWatermarkPredicateFunc: InternalRow =&gt; Boolean</p> <p>Predicate for late rows based on the <a href="#">stateWatermarkPredicate</a></p> <p>Used for the following:</p> <ul style="list-style-type: none"> <li>• <a href="#">storeAndJoinWithOtherSide</a> (and check out where to append a row to the <a href="#">SymmetricHashJoinStateM</a>)</li> <li>• <a href="#">removeOldState</a></li> </ul>

# JoinStateWatermarkPredicates — Watermark Predicates for State Removal

`JoinStateWatermarkPredicates` contains watermark predicates for state removal of the children of a `StreamingSymmetricHashJoinExec` physical operator:

- `JoinStateWatermarkPredicate` for the left-hand side of a join (default: `None`)
- `JoinStateWatermarkPredicate` for the right-hand side of a join (default: `None`)

`JoinStateWatermarkPredicates` is `created` for the following:

- `StreamingSymmetricHashJoinExec` physical operator is created (with the optional properties undefined, including `JoinStateWatermarkPredicates`)
- `StreamingSymmetricHashJoinHelper` utility is requested for `one` (for `IncrementalExecution` for the `state preparation rule` to optimize and specify the execution-specific configuration for a query plan with `StreamingSymmetricHashJoinExec` physical operators)

## Textual Representation — `toString` Method

`toString: String`

Note

`toString` is part of the `java.lang.Object` contract for the string representation of the object.

`toString` uses the `left` and `right` predicates for the string representation:

```
state cleanup [ left [left], right [right] ]
```

# JoinStateWatermarkPredicate Contract (Sealed Trait)

`JoinStateWatermarkPredicate` is the abstraction of join state watermark predicates that are described by a [Catalyst expression](#) and `desc`.

`JoinStateWatermarkPredicate` is created using [StreamingSymmetricHashJoinHelper](#) utility (for planning a [StreamingSymmetricHashJoinExec](#) physical operator for execution with execution-specific configuration)

`JoinStateWatermarkPredicate` is used to create a [OneSideHashJoiner](#) (and [JoinStateWatermarkPredicates](#)).

Table 1. JoinStateWatermarkPredicate Contract

Method	Description
<code>desc</code>	<code>desc: String</code> Used exclusively for the <a href="#">textual representation</a>
<code>expr</code>	<code>expr: Expression</code> A <a href="#">Catalyst Expression</a> Used for the <a href="#">textual representation</a> and a <a href="#">JoinStateWatermarkPredicates</a> (for <a href="#">StreamingSymmetricHashJoinExec</a> physical operator)

Table 2. JoinStateWatermarkPredicates

JoinStateWatermarkPredicate	Description
JoinStateKeyWatermarkPredicate	<p>Watermark predicate on state keys (i.e. when the <a href="#">streaming watermark</a> is defined either on the <a href="#">left</a> or <a href="#">right</a> join keys)</p> <p>Created when <a href="#">StreamingSymmetricHashJoinHelper</a> utility is requested for a <a href="#">JoinStateWatermarkPredicates</a> for the left and right side of a stream-stream join (when <a href="#">IncrementalExecution</a> is requested to optimize a query plan with a <a href="#">StreamingSymmetricHashJoinExec</a> physical operator)</p> <p>Used when <a href="#">OneSideHashJoiner</a> is requested for the <a href="#">stateKeyWatermarkPredicateFunc</a> and then to <a href="#">remove an old state</a></p>
JoinStateValueWatermarkPredicate	Watermark predicate on state values

Note	<code>JoinStateWatermarkPredicate</code> is a Scala <b>sealed trait</b> which means that all the <a href="#">implementations</a> are in the same compilation unit (a single file).
------	--

## Textual Representation — `toString` Method

```
toString: String
```

Note	<code>toString</code> is part of the <a href="#">java.lang.Object</a> contract for the string representation of the object.
------	---

`toString` uses the `desc` and `expr` for the string representation:

```
[desc]: [expr]
```

# StateStoreAwareZipPartitionsHelper — Extension Methods for Creating StateStoreAwareZipPartitionsRDD

`StateStoreAwareZipPartitionsHelper` is a **Scala implicit class** of a data RDD (of type `RDD[T]`) to [create a `StateStoreAwareZipPartitionsRDD`](#) for [StreamingSymmetricHashJoinExec](#) physical operator.

Note

[Implicit Classes](#) are a language feature in Scala for **implicit conversions** with **extension methods** for existing types.

## Creating StateStoreAwareZipPartitionsRDD — `stateStoreAwareZipPartitions` Method

```
stateStoreAwareZipPartitions[U: ClassTag, V: ClassTag](
  dataRDD2: RDD[U],
  stateInfo: StatefulOperatorStateInfo,
  storeNames: Seq[String],
  storeCoordinator: StateStoreCoordinatorRef
)(f: (Iterator[T], Iterator[U]) => Iterator[V]): RDD[V]
```

`stateStoreAwareZipPartitions` simply creates a new [StateStoreAwareZipPartitionsRDD](#).

Note

`stateStoreAwareZipPartitions` is used exclusively when [StreamingSymmetricHashJoinExec](#) physical operator is requested to [execute](#) and [generate a recipe for a distributed computation \(as an `RDD\[InternalRow\]`\)](#).

# StreamingSymmetricHashJoinHelper Utility

`StreamingSymmetricHashJoinHelper` is a Scala object with the following utility methods:

- `getStateWatermarkPredicates`

## Creating JoinStateWatermarkPredicates

### — `getStateWatermarkPredicates` Object Method

```
getStateWatermarkPredicates(  
    leftAttributes: Seq[Attribute],  
    rightAttributes: Seq[Attribute],  
    leftKeys: Seq[Expression],  
    rightKeys: Seq[Expression],  
    condition: Option[Expression],  
    eventTimeWatermark: Option[Long]): JoinStateWatermarkPredicates
```

`getStateWatermarkPredicates` tries to find the index of the [watermark attribute](#) among the left keys first, and if not found, the right keys.

Note	The <a href="#">watermark attribute</a> is defined using <code>Dataset.withWatermark</code> operator.
------	---

`getStateWatermarkPredicates` [determines the state watermark predicate](#) for the left side of a join (for the given `leftAttributes`, the `leftKeys` and the `rightAttributes`).

`getStateWatermarkPredicates` [determines the state watermark predicate](#) for the right side of a join (for the given `rightAttributes`, the `rightKeys` and the `leftAttributes`).

In the end, `getStateWatermarkPredicates` creates a [JoinStateWatermarkPredicates](#) with the left- and right-side state watermark predicates.

Note	<code>getStateWatermarkPredicates</code> is used exclusively when <code>IncrementalExecution</code> is requested to <a href="#">apply the state preparation rule for batch-specific configuration</a> (while optimizing query plans with <code>StreamingSymmetricHashJoinExec</code> physical operators).
------	---

## Join State Watermark Predicate (for One Side of Join)

### — `getOneSideStateWatermarkPredicate` Internal Method

```
getOneSideStateWatermarkPredicate(
  oneSideInputAttributes: Seq[Attribute],
  oneSideJoinKeys: Seq[Expression],
  otherSideInputAttributes: Seq[Attribute]): Option[JoinStateWatermarkPredicate]
```

`getOneSideStateWatermarkPredicate` finds what attributes were used to define the [watermark attribute](#) (the `oneSideInputAttributes` attributes, the [left or right join keys](#)) and creates a [JoinStateWatermarkPredicate](#) as follows:

- [JoinStateKeyWatermarkPredicate](#) if the watermark was defined on a join key (with the watermark expression for the index of the join key expression)
- [JoinStateValueWatermarkPredicate](#) if the watermark was defined among the `oneSideInputAttributes` (with the [state value watermark](#) based on the given `oneSideInputAttributes` and `otherSideInputAttributes` )

Note	<code>getOneSideStateWatermarkPredicate</code> creates no <a href="#">JoinStateWatermarkPredicate</a> ( <code>None</code> ) for no watermark found.
------	---

Note	<code>getStateWatermarkPredicates</code> is used exclusively to create a <a href="#">JoinStateWatermarkPredicates</a> .
------	---

# StreamingJoinHelper Utility

`StreamingJoinHelper` is a Scala object with the following utility methods:

- [getStateValueWatermark](#)

Tip

Enable `ALL` logging level for `org.apache.spark.sql.catalyst.analysis.StreamingJoinHelper` to see what happens inside.

Add the following line to `conf/log4j.properties`:

```
log4j.logger.org.apache.spark.sql.catalyst.analysis.StreamingJoinHelper=ALL
```

Refer to [Logging](#).

## State Value Watermark — `getStateValueWatermark` Object Method

```
getStateValueWatermark(  
    attributesToFindStateWatermarkFor: AttributeSet,  
    attributesWithEventWatermark: AttributeSet,  
    joinCondition: Option[Expression],  
    eventWatermark: Option[Long]): Option[Long]
```

`getStateValueWatermark` ...FIXME

Note

`getStateValueWatermark` is used when:

- `UnsupportedOperationChecker` utility is used to [checkForStreaming](#)
- `StreamingSymmetricHashJoinHelper` utility is used to [create a JoinStateWatermarkPredicates](#)

# Extending Structured Streaming with New Data Sources

Spark Structured Streaming uses Spark SQL for planning streaming queries (*preparing for execution*).

Spark SQL is migrating from the former Data Source API V1 to a new Data Source API V2, and so is Structured Streaming. That is exactly the reason for [BaseStreamingSource](#) and [BaseStreamingSink](#) APIs for the two different Data Source API's class hierarchies, for streaming sources and sinks, respectively.

Structured Streaming supports two [stream execution engines](#) (i.e. [Micro-Batch](#) and [Continuous](#)) with their own APIs.

[Micro-Batch Stream Processing](#) supports the old Data Source API V1 and the new modern Data Source API V2 with micro-batch-specific APIs for streaming sources and sinks.

[Continuous Stream Processing](#) supports the new modern Data Source API V2 only with continuous-specific APIs for streaming sources and sinks.

The following are the questions to think of (and answer) while considering development of a new data source for Structured Streaming. They are supposed to give you a sense of how much work and time it takes as well as what Spark version to support (e.g. 2.2 vs 2.4).

- Data Source API V1
- Data Source API V2
- [Micro-Batch Stream Processing \(Structured Streaming V1\)](#)
- [Continuous Stream Processing \(Structured Streaming V2\)](#)
- Read side ([BaseStreamingSource](#))
- Write side ([BaseStreamingSink](#))

# BaseStreamingSource Contract — Base of Streaming Readers and Sources

`BaseStreamingSource` is the abstraction of streaming readers and sources that can be stopped.

The main purpose of `BaseStreamingSource` is to share a common abstraction between the former Data Source API V1 ([Source API](#)) and the modern Data Source API V2 (until Spark Structured Streaming migrates to the Data Source API V2 fully).

Table 1. BaseStreamingSource Contract

Method	Description
<code>stop</code>	<pre>void stop()</pre> <p>Stops the streaming source or reader (and frees up any resources it may have allocated)</p> <p>Used when:</p> <ul style="list-style-type: none"> <li><code>StreamExecution</code> is requested to <a href="#">stop streaming sources and readers</a></li> <li><code>DataStreamReader</code> is requested to <a href="#">load data from a MicroBatchReadSupport data source</a> (for read schema)</li> </ul>

Table 2. BaseStreamingSources (Extensions Only)

BaseStreamingSource	Description
<code>ContinuousReader</code>	Data source readers in <a href="#">Continuous Stream Processing</a> (based on Data Source API V2)
<code>MemoryStreamBase</code>	Base implementation of <a href="#">ContinuousMemoryStream</a> (for Continuous Stream Processing) and <a href="#">MemoryStream</a> (for Micro-Batch Stream Processing)
<code>MicroBatchReader</code>	Data source readers in <a href="#">Micro-Batch Stream Processing</a> (based on Data Source API V2)
<code>Source</code>	Streaming sources for <a href="#">Micro-Batch Stream Processing</a> (based on Data Source API V1)



# BaseStreamingSink Contract — Base of Streaming Writers and Sinks

`BaseStreamingSink` is the abstraction of [streaming writers and sinks](#) with the only purpose of sharing a common abstraction between the former Data Source API V1 ([Sink API](#)) and the modern Data Source API V2 (until Spark Structured Streaming migrates to the Data Source API V2 fully).

`BaseStreamingSink` defines no methods.

Table 1. BaseStreamingSinks (Extensions Only)

<b>BaseStreamingSink</b>	<b>Description</b>
<code>MemorySinkBase</code>	Base contract for data sinks in <a href="#">memory data source</a>
<code>Sink</code>	Streaming sinks for <a href="#">Micro-Batch Stream Processing</a> (based on Data Source API V1)
<code>StreamWriteSupport</code>	Data source writers (based on Data Source API V2)

# StreamWriteSupport Contract — Writable Streaming Data Sources

`StreamWriteSupport` is the abstraction of `DataSourceV2` sinks that [create StreamWriters](#) for streaming write (when used in streaming queries in [MicroBatchExecution](#) and [ContinuousExecution](#)).

```
StreamWriter createStreamWriter(
    String queryId,
    StructType schema,
    OutputMode mode,
    DataSourceOptions options)
```

`createStreamWriter` creates a [StreamWriter](#) for streaming write and is used when the [stream execution thread for a streaming query](#) is [started](#) and requests the stream execution engines to start, i.e.

- `ContinuousExecution` is requested to [runContinuous](#)
- `MicroBatchExecution` is requested to [run a single streaming batch](#)

Table 1. StreamWriteSupports

StreamWriteSupport	Description
<a href="#">ConsoleSinkProvider</a>	Streaming sink for <code>console</code> data source format
<a href="#">ForeachWriterProvider</a>	
<a href="#">KafkaSourceProvider</a>	
<a href="#">MemorySinkV2</a>	

# StreamWriter Contract

`StreamWriter` is the [extension](#) of the `DataSourceWriter` contract to support epochs, i.e. [streaming writers](#) that can [abort](#) and [commit](#) writing jobs for a specified epoch.

Tip

Read up on [DataSourceWriter](#) in [The Internals of Spark SQL](#) book.

Table 1. StreamWriter Contract

Method	Description
abort	<pre>void abort(     long epochId,     WriterCommitMessage[] messages)</pre> <p>Aborts the writing job for a specified <code>epochId</code> and <code>WriterCommitMessages</code></p> <p>Used exclusively when <code>MicroBatchWriter</code> is requested to <a href="#">abort</a></p>
commit	<pre>void commit(     long epochId,     WriterCommitMessage[] messages)</pre> <p>Commits the writing job for a specified <code>epochId</code> and <code>WriterCommitMessages</code></p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>EpochCoordinator</code> is requested to <a href="#">commitEpoch</a></li> <li>• <code>MicroBatchWriter</code> is requested to <a href="#">commit</a></li> </ul>

Table 2. StreamWriters

StreamWriter	Description
<a href="#">ForeachWriterProvider</a>	<b>foreachWriter</b> data source
<a href="#">ConsoleWriter</a>	<b>console</b> data source
<a href="#">KafkaStreamWriter</a>	<b>kafka</b> data source
<a href="#">MemoryStreamWriter</a>	<b>memory</b> data source

# DataSource — Pluggable Data Provider Framework

Tip

Read up on [DataSource — Pluggable Data Provider Framework](#) in [The Internals of Spark SQL](#) online book.

## Creating DataSource Instance

`DataSource` takes the following to be created:

- `SparkSession`
- `className`, i.e. the fully-qualified class name or an alias of the data source
- Paths (default: `Nil`, i.e. an empty collection)
- Optional user-defined schema (default: `None`)
- Names of the partition columns (default: (empty))
- Optional `BucketSpec` (default: `None`)
- Configuration options (default: empty)
- Optional `CatalogTable` (default: `None`)

`DataSource` initializes the [internal properties](#).

## Generating Metadata of Streaming Source (Data Source API V1) — `sourceSchema` Internal Method

```
sourceSchema(): SourceInfo
```

`sourceSchema` creates a new instance of the [data source class](#) and branches off per the type, e.g. [StreamSourceProvider](#), [FileFormat](#) and [other types](#).

Note

`sourceSchema` is used exclusively when `DataSource` is requested for the [SourceInfo](#).

## StreamSourceProvider

For a [StreamSourceProvider](#), `sourceSchema` requests the `StreamSourceProvider` for the name and schema (of the [streaming source](#)).

In the end, `sourceSchema` returns the name and the schema as part of `SourceInfo` (with partition columns unspecified).

## FileFormat

For a `FileFormat`, `sourceSchema` ...FIXME

## Other Types

For any other data source type, `sourceSchema` simply throws an

`UnsupportedOperationException` :

```
Data source [className] does not support streamed reading
```

## Creating Streaming Source (Micro-Batch Stream Processing / Data Source API V1) — `createSource` Method

```
createSource(  
    metadataPath: String): Source
```

`createSource` creates a new instance of the [data source class](#) and branches off per the type, e.g. [StreamSourceProvider](#), [FileFormat](#) and [other types](#).

Note	<code>createSource</code> is used exclusively when <code>MicroBatchExecution</code> is requested to initialize the analyzed logical plan.
------	---

## StreamSourceProvider

For a [StreamSourceProvider](#), `createSource` requests the `StreamSourceProvider` to [create a source](#).

## FileFormat

For a `FileFormat`, `createSource` creates a new [FileStreamSource](#).

`createSource` throws an `IllegalArgumentException` when `path` option was not specified for a `FileFormat` data source:

```
'path' is not specified
```

## Other Types

For any other data source type, `createSource` simply throws an `UnsupportedOperationException`:

```
Data source [className] does not support streamed reading
```

## Creating Streaming Sink — `createSink` Method

```
createSink(  
    outputMode: OutputMode): Sink
```

`createSink` creates a [streaming sink](#) for [StreamSinkProvider](#) or [FileFormat](#) data sources.

Tip	Read up on <a href="#">FileFormat Data Source</a> in <a href="#">The Internals of Spark SQL</a> book.
-----	---

Internally, `createSink` creates a new instance of the [providingClass](#) and branches off per type:

- For a [StreamSinkProvider](#), `createSink` simply delegates the call and requests it to [create a streaming sink](#)
- For a [FileFormat](#), `createSink` creates a [FileStreamSink](#) when `path` option is specified and the output mode is [Append](#)

`createSink` throws a [IllegalArgumentException](#) when `path` option is not specified for a [FileFormat](#) data source:

```
'path' is not specified
```

`createSink` throws an [AnalysisException](#) when the given [OutputMode](#) is different from [Append](#) for a [FileFormat](#) data source:

```
Data source [className] does not support [outputMode] output mode
```

`createSink` throws an [UnsupportedOperationException](#) for unsupported data source formats:

```
Data source [className] does not support streamed writing
```

**Note**

`createSink` is used exclusively when `DataStreamWriter` is requested to [start a streaming query](#).

## Internal Properties

Name	Description
<code>providingClass</code>	<code>java.lang.Class</code> for the <a href="#">className</a> (that can be a fully-qualified class name or an alias of the data source)
<code>sourceInfo</code>	<p><code>sourceInfo: SourceInfo</code></p> <p>Metadata of a <a href="#">Source</a> with the alias (short name), the schema, and optional partitioning columns</p> <p><code>sourceInfo</code> is a lazy value and so initialized once (the very first time) when accessed.</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>DataSource</code> is requested to <a href="#">create a source (for a FileFormat data source)</a> (when <code>MicroBatchExecution</code> is requested to <a href="#">initialize the analyzed logical plan</a>)</li> <li>• <code>StreamingRelation</code> utility is requested for a <a href="#">StreamingRelation</a> (when <code>DataStreamReader</code> is requested for a <a href="#">streaming DataFrame</a>)</li> </ul>

# Demos

1. Demo: Internals of FlatMapGroupsWithStateExec Physical Operator
2. Demo: Exploring Checkpointed State
3. Demo: Streaming Watermark with Aggregation in Append Output Mode
4. Demo: Streaming Query for Running Counts (Socket Source and Complete Output Mode)
5. Demo: Streaming Aggregation with Kafka Data Source
6. Demo: groupByKey Streaming Aggregation in Update Mode
7. Demo: StateStoreSaveExec with Complete Output Mode
8. Demo: StateStoreSaveExec with Update Output Mode
9. Developing Custom Streaming Sink (and Monitoring SQL Queries in web UI)
10. current\_timestamp Function For Processing Time in Streaming Queries
11. Using StreamingQueryManager for Query Termination Management

# Demo: Internals of FlatMapGroupsWithStateExec Physical Operator

The following demo shows the internals of [FlatMapGroupsWithStateExec](#) physical operator in a [Arbitrary Stateful Streaming Aggregation](#).

```
// Reduce the number of partitions and hence the state stores
// That is supposed to make debugging state checkpointing easier
val numShufflePartitions = 1
import org.apache.spark.sql.internal.SQLConf.SHUFFLE_PARTITIONS
spark.sessionState.conf.setConf(SHUFFLE_PARTITIONS, numShufflePartitions)
assert(spark.sessionState.conf.numShufflePartitions == numShufflePartitions)

// Define event "format"
// Use :paste mode in spark-shell
import java.sql.Timestamp
case class Event(time: Timestamp, value: Long)
import scala.concurrent.duration._
object Event {
  def apply(secs: Long, value: Long): Event = {
    Event(new Timestamp(secs.seconds.toMillis), value)
  }
}

// Using memory data source for full control of the input
import org.apache.spark.sql.execution.streaming.MemoryStream
implicit val sqlCtx = spark.sqlContext
val events = MemoryStream[Event]
val values = events.toDS
assert(values.isStreaming, "values must be a streaming Dataset")

values.printSchema
/**
root
 |-- time: timestamp (nullable = true)
 |-- value: long (nullable = false)
 */

import scala.concurrent.duration._
val delayThreshold = 10.seconds
val valuesWatermarked = values
  .withWatermark(eventTime = "time", delayThreshold.toString) // required for EventTim
eTimeout

// Could use Long directly, but...
// Let's use case class to make the demo a bit more advanced
```



```

// Delete the checkpoint location from previous executions
import java.nio.file.{Files, FileSystems}
import java.util.Comparator
import scala.collection.JavaConverters._
val path = FileSystems.getDefault.getPath(checkpointLocation)
if (Files.exists(path)) {
  Files.walk(path)
    .sorted(Comparator.reverseOrder())
    .iterator
    .asScala
    .foreach(p => p.toFile.delete)
}

import org.apache.spark.sql.streaming.OutputMode.Update
val streamingQuery = valuesCounted
  .writeStream
  .format("memory")
  .queryName(queryName)
  .option("checkpointLocation", checkpointLocation)
  .outputMode(Update)
  .start

assert(streamingQuery.status.message == "Waiting for data to arrive")

// Use web UI to monitor the metrics of the streaming query
// Go to http://localhost:4040/SQL/ and click one of the Completed Queries with Job IDs

// You may also want to check out checkpointed state
// in /tmp/checkpoint-FlatMapGroupsWithStateExec_demo/state/0/0

val batch = Seq(
  Event(secs = 1, value = 1),
  Event(secs = 15, value = 2))
events.addData(batch)
streamingQuery.processAllAvailable()

/**
>>> keyCounts(key = 1, state = <empty>)
>>> >>> currentProcessingTimeMs: 1561881557237
>>> >>> currentWatermarkMs: 0
>>> >>> hasTimedOut: false
>>> keyCounts(key = 2, state = <empty>)
>>> >>> currentProcessingTimeMs: 1561881557237
>>> >>> currentWatermarkMs: 0
>>> >>> hasTimedOut: false
*/
spark.table(queryName).show(truncate = false)
/*
+-----+
|value|count|
+-----+

```

```

|1    | [1]  |
|2    | [1]  |
+----+----+
*/
// With at least one execution we can review the execution plan
streamingQuery.explain
/** 
== Physical Plan ==
*(2) Project [_1#928L AS value#931L, _2#929 AS count#932]
+- *(2) SerializeFromObject [assertnotnull(input[0, scala.Tuple2, true])._1 AS _1#928L
, if (isnull(assertnotnull(input[0, scala.Tuple2, true])._2)) null else named_struct(v
alue, assertnotnull(assertnotnull(input[0, scala.Tuple2, true])._2).value) AS _2#929]
   +- FlatMapGroupsWithState $line140.$read$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$Lambda$4117/181063008@d2cdc82, value#923: bigint, newInstance(class s
cala.Tuple2), [value#923L], [time#915-T10000ms, value#916L], obj#927: scala.Tuple2, st
ate info [ checkpoint = file:/tmp/checkpoint-FlatMapGroupsWithStateExec_demo/state, ru
nId = 95c3917c-2fd7-45b2-86f6-6c01f0115e1d, opId = 0, ver = 1, numPartitions = 1], cla
ss[value[0]: bigint], 2, Update, EventTimeTimeout, 1561881557499, 5000
   +- *(1) Sort [value#923L ASC NULLS FIRST], false, 0
      +- Exchange hashpartitioning(value#923L, 1)
         +- AppendColumns $line140.$read$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$Lambda$4118/2131767153@3e606b4c, newInstance(class scala.Tuple2), [in
put[0, bigint, false] AS value#923L]
         +- EventTimeWatermark time#915: timestamp, interval 10 seconds
            +- LocalTableScan <empty>, [time#915, value#916L]
*/
type Millis = Long
def toMillis(datetime: String): Millis = {
  import java.time.format.DateTimeFormatter
  import java.time.LocalDateTime
  import java.time.ZoneOffset
  LocalDateTime
    .parse(datetime, DateTimeFormatter.ISO_DATE_TIME)
    .toInstant(ZoneOffset.UTC)
    .toEpochMilli
}
val currentWatermark = streamingQuery.lastProgress.eventTime.get("watermark")
val currentWatermarkSecs = toMillis(currentWatermark).millis.toSeconds.seconds

val expectedWatermarkSecs = 5.seconds
assert(currentWatermarkSecs == expectedWatermarkSecs, s"Current event-time watermark i
s $currentWatermarkSecs, but should be $expectedWatermarkSecs (maximum event time - de
layThreshold ${delayThreshold.toMillis})")

// Let's access the FlatMapGroupsWithStateExec physical operator
import org.apache.spark.sql.execution.streaming.StreamingQueryWrapper
import org.apache.spark.sql.execution.streaming.StreamExecution
val engine: StreamExecution = streamingQuery

```

```

.asInstanceOf[StreamingQueryWrapper]
.streamingQuery

import org.apache.spark.sql.execution.streaming.IncrementalExecution
val lastMicroBatch: IncrementalExecution = engine.lastExecution

// Access executedPlan that is the optimized physical query plan ready for execution
// All streaming optimizations have been applied at this point
val plan = lastMicroBatch.executedPlan

// Find the FlatMapGroupsWithStateExec physical operator
import org.apache.spark.sql.execution.streaming.FlatMapGroupsWithStateExec
val flatMapOp = plan.collect { case op: FlatMapGroupsWithStateExec => op }.head

// Display metrics
import org.apache.spark.sql.execution.metric.SQLMetric
def formatMetrics(name: String, metric: SQLMetric) = {
  val desc = metric.name.getOrElse("")
  val value = metric.value
  f"$name%-30s | $desc%-69s | $value%-10s"
}
flatMapOp.metrics.map { case (name, metric) => formatMetrics(name, metric) }.foreach(println)
/*
| numTotalStateRows           | number of total state rows
| 0
| stateMemory                | memory used by state total (min, med, max)
| 390
| loadedMapCacheHitCount     | count of cache hit on states cache in provider
| 1
| numOutputRows               | number of output rows
| 0
| stateOnCurrentVersionSizeBytes | estimated size of state only on current version total (min, med, max) | 102
| loadedMapCacheMissCount    | count of cache miss on states cache in provider
| 0
| commitTimeMs                | time to commit changes total (min, med, max)
| -2
| allRemovalsTimeMs           | total time to remove rows total (min, med, max)
| -2
| numUpdatedStateRows          | number of updated state rows
| 0
| allUpdatesTimeMs             | total time to update rows total (min, med, max)
| -2
*/
val batch = Seq(
  Event(secs = 1, value = 1), // under the watermark (5000 ms) so it's disregarded
  Event(secs = 6, value = 3)) // above the watermark so it should be counted
events.addData(batch)
streamingQuery.processAllAvailable()

/*

```

```

>>> keyCounts(key = 3, state = <empty>)
>>> >>> currentProcessingTimeMs: 1561881643568
>>> >>> currentWatermarkMs: 5000
>>> >>> hasTimedOut: false
*/

```

```

spark.table(queryName).show(truncate = false)
/**+
+-----+-----+
|value|count|
+-----+-----+
|1    |[1]   |
|2    |[1]   |
|3    |[1]   |
+-----+-----+
*/

```

```

val batch = Seq(
  Event(secs = 17, value = 3)) // advances the watermark
events.addData(batch)
streamingQuery.processAllAvailable()

```

```

/**+
>>> keyCounts(key = 3, state = <empty>)
>>> >>> currentProcessingTimeMs: 1561881672887
>>> >>> currentWatermarkMs: 5000
>>> >>> hasTimedOut: false
*/

```

```

val currentWatermark = streamingQuery.lastProgress.eventTime.get("watermark")
val currentWatermarkSecs = toMillis(currentWatermark).millis.toSeconds.seconds

val expectedWatermarkSecs = 7.seconds
assert(currentWatermarkSecs == expectedWatermarkSecs, s"Current event-time watermark is $currentWatermarkSecs, but should be $expectedWatermarkSecs (maximum event time - delayThreshold ${delayThreshold.toMillis})")

```

```

spark.table(queryName).show(truncate = false)
/**+
+-----+-----+
|value|count|
+-----+-----+
|1    |[1]   |
|2    |[1]   |
|3    |[1]   |
|3    |[1]   |
+-----+-----+
*/

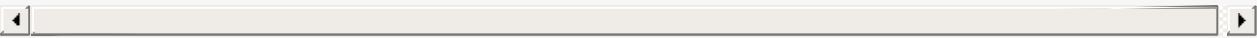
```

```

val batch = Seq(
  Event(secs = 18, value = 3)) // advances the watermark
events.addData(batch)
streamingQuery.processAllAvailable()

```

```
/**  
>>> keyCounts(key = 3, state = <empty>)  
>>> >>> currentProcessingTimeMs: 1561881778165  
>>> >>> currentWatermarkMs: 7000  
>>> >>> hasTimedOut: false  
*/  
  
// Eventually...  
streamingQuery.stop()
```



# Demo: Arbitrary Stateful Streaming Aggregation with KeyValueGroupedDataset.flatMapGroupsWithState Operator

The following demo shows an example of [Arbitrary Stateful Streaming Aggregation](#) with `KeyValueGroupedDataset.flatMapGroupsWithState` operator.

```
import java.sql.Timestamp
type DeviceId = Long
case class Signal(timestamp: Timestamp, deviceId: DeviceId, value: Long)

// input stream
import org.apache.spark.sql.functions._
val signals = spark
  .readStream
  .format("rate")
  .option("rowsPerSecond", 1)
  .load
  .withColumn("deviceId", rint(rand() * 10) cast "int") // 10 devices randomly assigned to values
  .withColumn("value", $"value" % 10) // randomize the values (just for fun)
  .as[Signal] // convert to our type (from "unpleasant" Row)

import org.apache.spark.sql.streaming.GroupState
type Key = Int
type Count = Long
type State = Map[Key, Count]
case class EventsCounted(deviceId: DeviceId, count: Long)
def countValuesPerDevice(
  deviceId: Int,
  signals: Iterator[Signal],
  state: GroupState[State]): Iterator[EventsCounted] = {
  val values = signals.toSeq
  println(s"Device: $deviceId")
  println(s"Signals (${values.size}):")
  values.zipWithIndex.foreach { case (v, idx) => println(s"$idx. $v") }
  println(s"State: $state")

  // update the state with the count of elements for the key
  val initialState: State = Map(deviceId -> 0)
  val oldState = state.getOption.getOrElse(initialState)
  // the name to highlight that the state is for the key only
  val newValue = oldState(deviceId) + values.size
  val newState = Map(deviceId -> newValue)
  state.update(newState)
}
```

## Arbitrary Stateful Streaming Aggregation with KeyValueGroupedDataset.flatMapGroupsWithState Operator

```
// you must not return as it's already consumed
// that leads to a very subtle error where no elements are in an iterator
// iterators are one-pass data structures
Iterator(EventsCounted(deviceId, newValue))
}

// stream processing using flatMapGroupsWithState operator
val deviceId: Signal => DeviceId = { case Signal(_, deviceId, _) => deviceId }
val signalsByDevice = signals.groupByKey(deviceId)

import org.apache.spark.sql.streaming.{GroupStateTimeout, OutputMode}
val signalCounter = signalsByDevice.flatMapGroupsWithState(
    outputMode = OutputMode.Append,
    timeoutConf = GroupStateTimeout.NoTimeout)(countValuesPerDevice)

import org.apache.spark.sql.streaming.{OutputMode, Trigger}
import scala.concurrent.duration._
val sq = signalCounter.
    writeStream.
    format("console").
    option("truncate", false).
    trigger(Trigger.ProcessingTime(10.seconds)).
    outputMode(OutputMode.Append).
    start
```

# Demo: Exploring Checkpointed State

The following demo shows the internals of the checkpointed state of a [stateful streaming query](#).

The demo uses the state checkpoint directory that was used in [Demo: Streaming Watermark with Aggregation in Append Output Mode](#).

```
// Change the path to match your configuration
val checkpointRootLocation = "/tmp/checkpoint-watermark_demo/state"
val version = 1L

import org.apache.spark.sql.execution.streaming.state.StateStoreId
val storeId = StateStoreId(
    checkpointRootLocation,
    operatorId = 0,
    partitionId = 0)

// The key and value schemas should match the watermark demo
// .groupBy(window($"time", windowDuration.toString) as "sliding_window")
import org.apache.spark.sql.types.{TimestampType, StructField, StructType}
val keySchema = StructType(
    StructField("sliding_window",
        StructType(
            StructField("start", TimestampType, nullable = true) :::
            StructField("end", TimestampType, nullable = true) :: Nil),
        nullable = false) :: Nil)
scala> keySchema.printTreeString
root
|-- sliding_window: struct (nullable = false)
|   |-- start: timestamp (nullable = true)
|   |-- end: timestamp (nullable = true)

// .agg(collect_list("batch") as "batches", collect_list("value") as "values")
import org.apache.spark.sql.types.{ArrayType, LongType}
val valueSchema = StructType(
    StructField("batches", ArrayType(LongType, true), true) :::
    StructField("values", ArrayType(LongType, true), true) :: Nil)
scala> valueSchema.printTreeString
root
|-- batches: array (nullable = true)
|   |-- element: long (containsNull = true)
|-- values: array (nullable = true)
|   |-- element: long (containsNull = true)

val indexOrdinal = None
import org.apache.spark.sql.execution.streaming.state.StateStoreConf
val storeConf = StateStoreConf(spark.sessionState.conf)
val hadoopConf = spark.sessionState.newHadoopConf()
```

```

import org.apache.spark.sql.execution.streaming.state.StateStoreProvider
val provider = StateStoreProvider.createAndInit(
  storeId, keySchema, valueSchema, indexOrdinal, storeConf, hadoopConf)

// You may want to use the following higher-level code instead
import java.util.UUID
val queryRunId = UUID.randomUUID
import org.apache.spark.sql.execution.streaming.state.StateStoreProviderId
val storeProviderId = StateStoreProviderId(storeId, queryRunId)
import org.apache.spark.sql.execution.streaming.state.StateStore
val store = StateStore.get(
  storeProviderId,
  keySchema,
  valueSchema,
  indexOrdinal,
  version,
  storeConf,
  hadoopConf)

import org.apache.spark.sql.execution.streaming.state.UnsafeRowPair
def formatRowPair(rowPair: UnsafeRowPair) = {
  s"${rowPair.key.getLong(0)}, ${rowPair.value.getLong(0)}"
}
store.iterator.map(formatRowPair).foreach(println)

// WIP: Missing value (per window)
def formatRowPair(rowPair: UnsafeRowPair) = {
  val window = rowPair.key.getStruct(0, 2)
  import scala.concurrent.duration._
  val begin = window.getLong(0).millis.toSeconds
  val end = window.getLong(1).millis.toSeconds

  val value = rowPair.value.getStruct(0, 4)
  // input is (time, value, batch) all longs
  val t = value.getLong(1).millis.toSeconds
  val v = value.getLong(2)
  val b = value.getLong(3)
  s"(key: [$begin, $end], ($t, $v, $b))"
}
store.iterator.map(formatRowPair).foreach(println)

```

# Demo: Streaming Watermark with Aggregation in Append Output Mode

The following demo shows the behaviour and the internals of [streaming watermark](#) with a [streaming aggregation](#) in [Append](#) output mode.

The demo also shows the behaviour and the internals of [StateStoreSaveExec](#) physical operator in [Append output mode](#).

Tip

The below code is part of [StreamingAggregationAppendMode](#) streaming application.

```
// Reduce the number of partitions and hence the state stores
// That is supposed to make debugging state checkpointing easier
val numShufflePartitions = 1
import org.apache.spark.sql.internal.SQLConf.SHUFFLE_PARTITIONS
spark.sessionState.conf.setConf(SHUFFLE_PARTITIONS, numShufflePartitions)
assert(spark.sessionState.conf.numShufflePartitions == numShufflePartitions)

// Define event "format"
// Use :paste mode in spark-shell
import java.sql.Timestamp
case class Event(time: Timestamp, value: Long, batch: Long)
import scala.concurrent.duration._
object Event {
  def apply(secs: Long, value: Long, batch: Long): Event = {
    Event(new Timestamp(secs.seconds.toMillis), value, batch)
  }
}

// Using memory data source for full control of the input
import org.apache.spark.sql.execution.streaming.MemoryStream
implicit val sqlCtx = spark.sqlContext
val events = MemoryStream[Event]
val values = events.toDS
assert(values.isStreaming, "values must be a streaming Dataset")

values.printSchema
/*
root
 |-- time: timestamp (nullable = true)
 |-- value: long (nullable = false)
 |-- batch: long (nullable = false)
 */

// Streaming aggregation using groupBy operator to demo StateStoreSaveExec operator
// Define required watermark for late events for Append output mode
```

```

import scala.concurrent.duration._
val delayThreshold = 10.seconds
val eventTime = "time"

val valuesWatermarked = values
    .withWatermark(eventTime, delayThreshold.toString) // defines watermark (before groupBy!)

// EventTimeWatermark logical operator is planned as EventTimeWatermarkExec physical operator
// Note that as a physical operator EventTimeWatermarkExec shows itself without the Exec suffix
valuesWatermarked.explain
/***
== Physical Plan ==
EventTimeWatermark time#3: timestamp, interval 10 seconds
+- StreamingRelation MemoryStream[time#3,value#4L,batch#5L], [time#3, value#4L, batch#5L]
*/

```

```

val windowDuration = 5.seconds
import org.apache.spark.sql.functions.window
val countsPer5secWindow = valuesWatermarked
    .groupBy(window(col(eventTime), windowDuration.toString) as "sliding_window")
    .agg(collect_list("batch") as "batches", collect_list("value") as "values")

countsPer5secWindow.printSchema
/***
root
 |-- sliding_window: struct (nullable = false)
 |   |-- start: timestamp (nullable = true)
 |   |-- end: timestamp (nullable = true)
 |-- batches: array (nullable = true)
 |   |-- element: long (containsNull = true)
 |-- values: array (nullable = true)
 |   |-- element: long (containsNull = true)
*/

```

```

// valuesPerGroupWindowed is a streaming Dataset with just one source
// It knows nothing about output mode or watermark yet
// That's why StatefulOperatorStateInfo is generic
// and no batch-specific values are printed out
// That will be available after the first streaming batch
// Use sq.explain to know the runtime-specific values
countsPer5secWindow.explain
/***
== Physical Plan ==
ObjectHashAggregate(keys=[window#23-T10000ms], functions=[collect_list(batch#5L, 0, 0)
, collect_list(value#4L, 0, 0)])
+- StateStoreSave [window#23-T10000ms], state info [ checkpoint = <unknown>, runId = 5
0e62943-fe5d-4a02-8498-7134ecbf5122, opId = 0, ver = 0, numPartitions = 1], Append, 0,
2
+- ObjectHashAggregate(keys=[window#23-T10000ms], functions=[merge_collect_list(bat

```

```

ch#5L, 0, 0), merge_collect_list(value#4L, 0, 0)])
    +- StateStoreRestore [window#23-T10000ms], state info [ checkpoint = <unknown>,
runId = 50e62943-fe5d-4a02-8498-7134ecbf5122, opId = 0, ver = 0, numPartitions = 1], 2
        +- ObjectHashAggregate(keys=[window#23-T10000ms], functions=[merge_collect_li
st(batch#5L, 0, 0), merge_collect_list(value#4L, 0, 0)])
            +- Exchange hashpartitioning(window#23-T10000ms, 1)
                +- ObjectHashAggregate(keys=[window#23-T10000ms], functions=[partial_co
llect_list(batch#5L, 0, 0), partial_collect_list(value#4L, 0, 0)])
                    +- *(1) Project [named_struct(start, precisetimestampconversion(((((
CASE WHEN (cast(CEIL((cast((precisetimestampconversion(time#3-T10000ms, TimestampType,
LongType) - 0) as double) / 5000000.0)) as double) = (cast((precisetimestampconversio
n(time#3-T10000ms, TimestampType, LongType) - 0) as double) / 5000000.0)) THEN (CEIL(((
cast((precisetimestampconversion(time#3-T10000ms, TimestampType, LongType) - 0) as dou
ble) / 5000000.0)) + 1) ELSE CEIL((cast((precisetimestampconversion(time#3-T10000ms, T
imestampType, LongType) - 0) as double) / 5000000.0)) END + 0) - 1) * 5000000) + 0), L
ongType, TimestampType), end, precisetimestampconversion((((CASE WHEN (cast(CEIL((cas
t((precisetimestampconversion(time#3-T10000ms, TimestampType, LongType) - 0) as double
) / 5000000.0)) as double) = (cast((precisetimestampconversion(time#3-T10000ms, Timest
ampType, LongType) - 0) as double) / 5000000.0)) THEN (CEIL((cast((precisetimestampcon
version(time#3-T10000ms, TimestampType, LongType) - 0) as double) / 5000000.0)) + 1) E
LSE CEIL((cast((precisetimestampconversion(time#3-T10000ms, TimestampType, LongType) -
0) as double) / 5000000.0)) END + 0) - 1) * 5000000) + 5000000), LongType, TimestampT
ype)) AS window#23-T10000ms, value#4L, batch#5L]
                    +- *(1) Filter isnotnull(time#3-T10000ms)
                        +- EventTimeWatermark time#3: timestamp, interval 10 seconds
                            +- StreamingRelation MemoryStream[time#3,value#4L,batch#5L]
, [time#3, value#4L, batch#5L]
*/

```

```

val queryName = "watermark_demo"
val checkpointLocation = s"/tmp/checkpoint-$queryName"

// Delete the checkpoint location from previous executions
import java.nio.file.{Files, FileSystems}
import java.util.Comparator
import scala.collection.JavaConverters._
val path = FileSystems.getDefault.getPath(checkpointLocation)
if (Files.exists(path)) {
  Files.walk(path)
    .sorted(Comparator.reverseOrder())
    .iterator
    .asScala
    .foreach(p => p.toFile.delete)
}

// FIXME Use foreachBatch for batchId and the output Dataset
// Start the query and hence StateStoreSaveExec
import scala.concurrent.duration._
import org.apache.spark.sql.streaming.OutputMode
val streamingQuery = countsPer5secWindow
  .writeStream
  .format("memory")
  .queryName(queryName)

```

```

.option("checkpointLocation", checkpointLocation)
.outputMode(OutputMode.Append) // <-- Use Append output mode
.start

assert(streamingQuery.status.message == "Waiting for data to arrive")

type Millis = Long
def toMillis(datetime: String): Millis = {
  import java.time.format.DateTimeFormatter
  import java.time.LocalDateTime
  import java.time.ZoneOffset
  LocalDateTime
    .parse(datetime, DateTimeFormatter.ISO_DATE_TIME)
    .toInstant(ZoneOffset.UTC)
    .toEpochMilli
}

// Use web UI to monitor the state of state (no pun intended)
// StateStoreSave and StateStoreRestore operators all have state metrics
// Go to http://localhost:4040/SQL/ and click one of the Completed Queries with Job IDs

// You may also want to check out checkpointed state
// in /tmp/checkpoint-watermark_demo/state/0/0

// The demo is aimed to show the following:
// 1. The current watermark
// 2. Check out the stats:
// - expired state (below the current watermark, goes to output and purged later)
// - late state (dropped as if never received and processed)
// - saved state rows (above the current watermark)

val batch = Seq(
  Event(1, 1, batch = 1),
  Event(15, 2, batch = 1))
events.addData(batch)
streamingQuery.processAllAvailable()

println(streamingQuery.lastProgress.stateOperators(0).prettyJson)
/***
{
  "numRowsTotal" : 1,
  "numRowsUpdated" : 0,
  "memoryUsedBytes" : 1102,
  "customMetrics" : {
    "loadedMapCacheHitCount" : 2,
    "loadedMapCacheMissCount" : 0,
    "stateOnCurrentVersionSizeBytes" : 414
  }
}
*/
val currentWatermark = streamingQuery.lastProgress.eventTime.get("watermark")

```

```

val currentWatermarkMs = toMillis(currentWatermark)

val maxTime = batch.maxBy(_.time.toInstant.toEpochMilli).time.toInstant.toEpochMilli.millis.toSeconds
val expectedMaxTime = 15
assert(maxTime == expectedMaxTime, s"Maximum time across events per batch is $maxTime, but should be $expectedMaxTime")

val expectedWatermarkMs = 5.seconds.toMillis
assert(currentWatermarkMs == expectedWatermarkMs, s"Current event-time watermark is ${currentWatermarkMs}, but should be $expectedWatermarkMs (maximum event time ${maxTime.seconds.toMillis} minus delayThreshold ${delayThreshold.toMillis})")

// FIXME Saved State Rows
// Use the metrics of the StateStoreSave operator
// Or simply streamingQuery.lastProgress.stateOperators.head
spark.table(queryName).orderBy("sliding_window").show(truncate = false)
/**
+-----+-----+
|sliding_window          |batches|values|
+-----+-----+
|[1970-01-01 01:00:00, 1970-01-01 01:00:05]| [1]    | [1]   |
+-----+-----+
*/
// With at least one execution we can review the execution plan
streamingQuery.explain
/**
scala> streamingQuery.explain
== Physical Plan ==
ObjectHashAggregate(keys=[window#18-T10000ms], functions=[collect_list(batch#5L, 0, 0)
, collect_list(value#4L, 0, 0)])
+- StateStoreSave [window#18-T10000ms], state info [ checkpoint = file:/tmp/checkpoint
-watermark_demo/state, runId = 73bb0ede-20f2-400d-8003-aa2fbebdd2e1, opId = 0, ver = 1
, numPartitions = 1], Append, 5000, 2
  +- ObjectHashAggregate(keys=[window#18-T10000ms], functions=[merge_collect_list(batch#5L, 0, 0), merge_collect_list(value#4L, 0, 0)])
    +- StateStoreRestore [window#18-T10000ms], state info [ checkpoint = file:/tmp/c
heckpoint-watermark_demo/state, runId = 73bb0ede-20f2-400d-8003-aa2fbebdd2e1, opId = 0
, ver = 1, numPartitions = 1], 2
      +- ObjectHashAggregate(keys=[window#18-T10000ms], functions=[merge_collect_li
st(batch#5L, 0, 0), merge_collect_list(value#4L, 0, 0)])
        +- Exchange hashpartitioning(window#18-T10000ms, 1)
          +- ObjectHashAggregate(keys=[window#18-T10000ms], functions=[partial_co
llect_list(batch#5L, 0, 0), partial_collect_list(value#4L, 0, 0)])
            +- *(1) Project [named_struct(start, precisetimestampconversion((((
CASE WHEN (cast(CEIL((cast((precisetimestampconversion(time#3-T10000ms, TimestampType,
LongType) - 0) as double) / 5000000.0)) as double) = (cast((precisetimestampconversion(
time#3-T10000ms, TimestampType, LongType) - 0) as double) / 5000000.0)) THEN (CEIL((
cast((precisetimestampconversion(time#3-T10000ms, TimestampType, LongType) - 0) as dou
ble) / 5000000.0)) + 1) ELSE CEIL((cast((precisetimestampconversion(time#3-T10000ms, T
imestampType, LongType) - 0) as double) / 5000000.0)) END + 0) - 1) * 5000000) + 0), L
ongType, TimestampType), end, precisetimestampconversion((((CASE WHEN (cast(CEIL((cas

```

```

t((precisetimestampconversion(time#3-T10000ms, TimestampType, LongType) - 0) as double
 ) / 5000000.0)) as double) = (cast((precisetimestampconversion(time#3-T10000ms, Timest
ampType, LongType) - 0) as double) / 5000000.0)) THEN (CEIL((cast((precisetimestampcon
version(time#3-T10000ms, TimestampType, LongType) - 0) as double) / 5000000.0)) + 1) E
LSE CEIL((cast((precisetimestampconversion(time#3-T10000ms, TimestampType, LongType)
 - 0) as double) / 5000000.0)) END + 0) - 1) * 5000000) + 5000000), LongType, TimestampT
ype)) AS window#18-T10000ms, value#4L, batch#5L]
      +- *(1) Filter isnotnull(time#3-T10000ms)
      +- EventTimeWatermark time#3: timestamp, interval 10 seconds
         +- LocalTableScan <empty>, [time#3, value#4L, batch#5L]
*/

```

```

import org.apache.spark.sql.execution.streaming.StreamingQueryWrapper
val engine = streamingQuery
  .asInstanceOf[StreamingQueryWrapper]
  .streamingQuery
import org.apache.spark.sql.execution.streaming.StreamExecution
assert(engine.isInstanceOf[StreamExecution])

val lastMicroBatch = engine.lastExecution
import org.apache.spark.sql.execution.streaming.IncrementalExecution
assert(lastMicroBatch.isInstanceOf[IncrementalExecution])

// Access executedPlan that is the optimized physical query plan ready for execution
// All streaming optimizations have been applied at this point
// We just need the EventTimeWatermarkExec physical operator
val plan = lastMicroBatch.executedPlan

// Let's find the EventTimeWatermarkExec physical operator in the plan
// There should be one only
import org.apache.spark.sql.execution.streaming.EventTimeWatermarkExec
val watermarkOp = plan.collect { case op: EventTimeWatermarkExec => op }.head

// Let's check out the event-time watermark stats
// They correspond to the concrete EventTimeWatermarkExec operator for a micro-batch
val stats = watermarkOp.eventTimeStats.value
import org.apache.spark.sql.execution.streaming.EventTimeStats
assert(stats.isInstanceOf[EventTimeStats])

println(stats)
/**
EventTimeStats(-9223372036854775808, 9223372036854775807, 0.0, 0)
*/

val batch = Seq(
  Event(1, 1, batch = 2),
  Event(15, 2, batch = 2),
  Event(35, 3, batch = 2))
events.addData(batch)
streamingQuery.processAllAvailable()

val currentWatermark = streamingQuery.lastProgress.eventTime.get("watermark")
val currentWatermarkMs = toMillis(currentWatermark)

```

```

val maxTime = batch.maxBy(_.time.toInstant.toEpochMilli).time.toInstant.toEpochMilli.millis.toSeconds
val expectedMaxTime = 35
assert(maxTime == expectedMaxTime, s"Maximum time across events per batch is $maxTime, but should be $expectedMaxTime")

val expectedWatermarkMs = 25.seconds.toMillis
assert(currentWatermarkMs == expectedWatermarkMs, s"Current event-time watermark is $currentWatermarkMs, but should be $expectedWatermarkMs (maximum event time ${maxTime.seconds.toMillis} minus delayThreshold ${delayThreshold.toMillis})")

// FIXME Expired State
// FIXME Late Events
// FIXME Saved State Rows
spark.table(queryName).orderBy("sliding_window").show(truncate = false)
/*
+-----+-----+
|sliding_window          |batches|values|
+-----+-----+
|[1970-01-01 01:00:00, 1970-01-01 01:00:05]| [1] | [1] |
|[1970-01-01 01:00:15, 1970-01-01 01:00:20]| [1, 2] | [2, 2] |
+-----+-----+
*/
// Check out the event-time watermark stats
val plan = engine.lastExecution.executedPlan
import org.apache.spark.sql.execution.streaming.EventTimeWatermarkExec
val watermarkOp = plan.collect { case op: EventTimeWatermarkExec => op }.head
val stats = watermarkOp.eventTimeStats.value
import org.apache.spark.sql.execution.streaming.EventTimeStats
assert(stats.isInstanceOf[EventTimeStats])

println(stats)
/*
EventTimeStats(-9223372036854775808, 9223372036854775807, 0.0, 0)
*/

val batch = Seq(
  Event(15, 1, batch = 3),
  Event(15, 2, batch = 3),
  Event(20, 3, batch = 3),
  Event(26, 4, batch = 3))
events.addData(batch)
streamingQuery.processAllAvailable()

val currentWatermark = streamingQuery.lastProgress.eventTime.get("watermark")
val currentWatermarkMs = toMillis(currentWatermark)

val maxTime = batch.maxBy(_.time.toInstant.toEpochMilli).time.toInstant.toEpochMilli.millis.toSeconds
val expectedMaxTime = 26
assert(maxTime == expectedMaxTime, s"Maximum time across events per batch is $maxTime,

```

```

        but should be $expectedMaxTime")

// Current event-time watermark should be the same as previously
// val expectedWatermarkMs = 25.seconds.toMillis
// The current max time is merely 26 so subtracting delayThreshold gives merely 16
assert(currentWatermarkMs == expectedWatermarkMs, s"Current event-time watermark is ${currentWatermarkMs}, but should be $expectedWatermarkMs (maximum event time ${maxTime.seconds.toMillis} minus delayThreshold ${delayThreshold.toMillis})")

// FIXME Expired State
// FIXME Late Events
// FIXME Saved State Rows
spark.table(queryName).orderBy("sliding_window").show(truncate = false)
/*
+-----+-----+
|sliding_window          |batches|values|
+-----+-----+
|[1970-01-01 01:00:00, 1970-01-01 01:00:05]| [1] | [1] |
|[1970-01-01 01:00:15, 1970-01-01 01:00:20]| [1, 2] | [2, 2] |
+-----+-----+
*/
// Check out the event-time watermark stats
val plan = engine.lastExecution.executedPlan
import org.apache.spark.sql.execution.streaming.EventTimeWatermarkExec
val watermarkOp = plan.collect { case op: EventTimeWatermarkExec => op }.head
val stats = watermarkOp.eventTimeStats.value
import org.apache.spark.sql.execution.streaming.EventTimeStats
assert(stats.isInstanceOf[EventTimeStats])

println(stats)
/*
EventTimeStats(26000,15000,19000.0,4)
*/

val batch = Seq(
  Event(36, 1, batch = 4))
events.addData(batch)
streamingQuery.processAllAvailable()

val currentWatermark = streamingQuery.lastProgress.eventTime.get("watermark")
val currentWatermarkMs = toMillis(currentWatermark)

val maxTime = batch.maxBy(_.time.toInstant.toEpochMilli).time.toInstant.toEpochMilli.millis.toSeconds
val expectedMaxTime = 36
assert(maxTime == expectedMaxTime, s"Maximum time across events per batch is $maxTime, but should be $expectedMaxTime")

val expectedWatermarkMs = 26.seconds.toMillis
assert(currentWatermarkMs == expectedWatermarkMs, s"Current event-time watermark is ${currentWatermarkMs}, but should be $expectedWatermarkMs (maximum event time ${maxTime.seconds.toMillis} minus delayThreshold ${delayThreshold.toMillis})")

```

```

// FIXME Expired State
// FIXME Late Events
// FIXME Saved State Rows
spark.table(queryName).orderBy("sliding_window").show(truncate = false)
/*
+-----+-----+
|sliding_window|batches|values|
+-----+-----+
|[1970-01-01 01:00:00, 1970-01-01 01:00:05]| [1] | [1] |
|[1970-01-01 01:00:15, 1970-01-01 01:00:20]| [1, 2] | [2, 2] |
+-----+-----+
*/

// Check out the event-time watermark stats
val plan = engine.lastExecution.executedPlan
import org.apache.spark.sql.execution.streaming.EventTimeWatermarkExec
val watermarkOp = plan.collect { case op: EventTimeWatermarkExec => op }.head
val stats = watermarkOp.eventTimeStats.value
import org.apache.spark.sql.execution.streaming.EventTimeStats
assert(stats.isInstanceOf[EventTimeStats])

println(stats)
/*
EventTimeStats(-9223372036854775808, 9223372036854775807, 0.0, 0)
*/

val batch = Seq(
  Event(50, 1, batch = 5)
)
events.addData(batch)
streamingQuery.processAllAvailable()

val currentWatermark = streamingQuery.lastProgress.eventTime.get("watermark")
val currentWatermarkMs = toMillis(currentWatermark)

val maxTime = batch.maxBy(_.time.toInstant.toEpochMilli).time.toInstant.toEpochMilli.millis.toSeconds
val expectedMaxTime = 50
assert(maxTime == expectedMaxTime, s"Maximum time across events per batch is $maxTime, but should be $expectedMaxTime")

val expectedWatermarkMs = 40.seconds.toMillis
assert(currentWatermarkMs == expectedWatermarkMs, s"Current event-time watermark is ${currentWatermarkMs}, but should be $expectedWatermarkMs (maximum event time ${maxTime.seconds.toMillis} minus delayThreshold ${delayThreshold.toMillis})")

// FIXME Expired State
// FIXME Late Events
// FIXME Saved State Rows
spark.table(queryName).orderBy("sliding_window").show(truncate = false)
/*
+-----+-----+

```

```
|sliding_window |batches|values|
+-----+-----+
|[1970-01-01 01:00:00, 1970-01-01 01:00:05]| [1] |[1]   |
|[1970-01-01 01:00:15, 1970-01-01 01:00:20]| [1, 2] |[2, 2]|
|[1970-01-01 01:00:25, 1970-01-01 01:00:30]| [3] |[4]   |
|[1970-01-01 01:00:35, 1970-01-01 01:00:40]| [2, 4] |[3, 1]|
+-----+-----+
*/
// Check out the event-time watermark stats
val plan = engine.lastExecution.executedPlan
import org.apache.spark.sql.execution.streaming.EventTimeWatermarkExec
val watermarkOp = plan.collect { case op: EventTimeWatermarkExec => op }.head
val stats = watermarkOp.eventTimeStats.value
import org.apache.spark.sql.execution.streaming.EventTimeStats
assert(stats.isInstanceOf[EventTimeStats])

println(stats)
/**
EventTimeStats(-9223372036854775808, 9223372036854775807, 0.0, 0)
*/
// Eventually...
streamingQuery.stop()
```

# Demo: Streaming Query for Running Counts (Socket Source and Complete Output Mode)

The following code shows a [streaming aggregation](#) (with `Dataset.groupBy` operator) in [complete](#) output mode that reads text lines from a socket (using socket data source) and outputs running counts of the words.

**Note**

The example is "borrowed" from [the official documentation of Spark](#). Changes and errors are only mine.

**Important**

Run `nc -lk 9999` first before running the demo.

```
// START: Only for easier debugging
// Reduce the number of partitions
// The state is then only for one partition
// which should make monitoring easier
val numShufflePartitions = 1
import org.apache.spark.sql.internal.SQLConf.SHUFFLE_PARTITIONS
spark.sessionState.conf.setConf(SHUFFLE_PARTITIONS, numShufflePartitions)

assert(spark.sessionState.conf.numShufflePartitions == numShufflePartitions)
// END: Only for easier debugging

val lines = spark
  .readStream
  .format("socket")
  .option("host", "localhost")
  .option("port", 9999)
  .load

scala> lines.printSchema
root
 |-- value: string (nullable = true)

import org.apache.spark.sql.functions.explode
val words = lines
  .select(explode(split($"value", """\W+""")) as "word")

val counts = words.groupBy("word").count

scala> counts.printSchema
root
 |-- word: string (nullable = true)
 |-- count: long (nullable = false)

// nc -lk 9999 is supposed to be up at this point
```

```

val queryName = "running_counts"
val checkpointLocation = s"/tmp/checkpoint-$queryName"

// Delete the checkpoint location from previous executions
import java.nio.file.{Files, FileSystems}
import java.util.Comparator
import scala.collection.JavaConverters._
val path = FileSystems.getDefault.getPath(checkpointLocation)
if (Files.exists(path)) {
  Files.walk(path)
    .sorted(Comparator.reverseOrder())
    .iterator
    .asScala
    .foreach(p => p.toFile.delete)
}

import org.apache.spark.sql.streaming.OutputMode.Complete
val runningCounts = counts
  .writeStream
  .format("console")
  .option("checkpointLocation", checkpointLocation)
  .outputMode(Complete)
  .start

scala> runningCounts.explain
== Physical Plan ==
WriteToDataSourceV2 org.apache.spark.sql.execution.streaming.sources.MicroBatchWriter@
205f195c
+- *(5) HashAggregate(keys=[word#72], functions=[count(1)])
  +- StateStoreSave [word#72], state info [ checkpoint = file:/tmp/checkpoint-running
  _counts/state, runId = f3b2e642-1790-4a17-ab61-3d894110b063, opId = 0, ver = 0, numPar
  titions = 1], Complete, 0, 2
    +- *(4) HashAggregate(keys=[word#72], functions=[merge_count(1)])
      +- StateStoreRestore [word#72], state info [ checkpoint = file:/tmp/checkpoin
t-running_counts/state, runId = f3b2e642-1790-4a17-ab61-3d894110b063, opId = 0, ver = 0
, numPartitions = 1], 2
        +- *(3) HashAggregate(keys=[word#72], functions=[merge_count(1)])
          +- Exchange hashpartitioning(word#72, 1)
            +- *(2) HashAggregate(keys=[word#72], functions=[partial_count(1)])
              +- Generate explode(split(value#83, \W+)), false, [word#72]
                +- *(1) Project [value#83]
                  +- *(1) ScanV2 socket[value#83] (Options: [host=localhost,p
ort=9999])

// Type lines (words) in the terminal with nc
// Observe the counts in spark-shell

// Use web UI to monitor the state of state (no pun intended)
// StateStoreSave and StateStoreRestore operators all have state metrics
// Go to http://localhost:4040/SQL/ and click one of the Completed Queries with Job IDs

// You may also want to check out checkpointed state

```

```
// in /tmp/checkpoint-running_counts/state/0/0  
  
// Eventually...  
runningCounts.stop()
```



# Demo: Streaming Aggregation with Kafka Data Source

The following example code shows a streaming aggregation (with [Dataset.groupBy](#) operator) that reads records from Kafka (with [Kafka Data Source](#)).

**Important**

Start up Kafka cluster and spark-shell with `spark-sql-kafka-0-10` package before running the demo.

**Tip**

You may want to consider copying the following code to `append.txt` and using `:load append.txt` command in spark-shell to load it (rather than copying and pasting it).

```
// START: Only for easier debugging
// The state is then only for one partition
// which should make monitoring easier
val numShufflePartitions = 1
import org.apache.spark.sql.internal.SQLConf.SHUFFLE_PARTITIONS
spark.sessionState.conf.setConf(SHUFFLE_PARTITIONS, numShufflePartitions)

assert(spark.sessionState.conf.numShufflePartitions == numShufflePartitions)
// END: Only for easier debugging

val records = spark
  .readStream
  .format("kafka")
  .option("subscribePattern", """topic-\d{2}""") // topics with two digits at the end
  .option("kafka.bootstrap.servers", ":9092")
  .load
scala> records.printSchema
root
|-- key: binary (nullable = true)
|-- value: binary (nullable = true)
|-- topic: string (nullable = true)
|-- partition: integer (nullable = true)
|-- offset: long (nullable = true)
|-- timestamp: timestamp (nullable = true)
|-- timestampType: integer (nullable = true)

// Since the streaming query uses Append output mode
// it has to define a streaming event-time watermark (using Dataset.withWatermark operator)
// UnsupportedOperationChecker makes sure that the requirement holds
val ids = records
  .withColumn("tokens", split($"value", ","))
  .withColumn("seconds", 'tokens(0) cast "long")
  .withColumn("event_time", to_timestamp(from_unixtime('seconds))) // <-- Event time h
```

```

as to be a timestamp
    .withColumn("id", 'tokens(1))
    .withColumn("batch", 'tokens(2) cast "int")
    .withWatermark(eventTime = "event_time", delayThreshold = "10 seconds") // <-- define watermark (before groupBy!)
    .groupBy($"event_time") // <-- use event_time for grouping
    .agg(collect_list("batch") as "batches", collect_list("id") as "ids")
    .withColumn("event_time", to_timestamp($"event_time")) // <-- convert to human-readable date
scala> ids.printSchema
root
|-- event_time: timestamp (nullable = true)
|-- batches: array (nullable = true)
|   |-- element: integer (containsNull = true)
|-- ids: array (nullable = true)
|   |-- element: string (containsNull = true)

assert(ids.isStreaming, "ids is a streaming query")

// ids knows nothing about the output mode or the current streaming watermark yet
// - Output mode is defined on writing side
// - streaming watermark is read from rows at runtime
// That's why StatefulOperatorStateInfo is generic (and uses the default Append for output mode)
// and no batch-specific values are printed out
// They will be available right after the first streaming batch
// Use explain on a streaming query to know the trigger-specific values
scala> ids.explain
== Physical Plan ==
ObjectHashAggregate(keys=[event_time#118-T10000ms], functions=[collect_list(batch#141,
0, 0), collect_list(id#129, 0, 0)])
+- StateStoreSave [event_time#118-T10000ms], state info [ checkpoint = <unknown>, runId = a870e6e2-b925-4104-9886-b211c0be1b73, opId = 0, ver = 0, numPartitions = 1], Append
, 0, 2
    +- ObjectHashAggregate(keys=[event_time#118-T10000ms], functions=[merge_collect_list(batch#141, 0, 0), merge_collect_list(id#129, 0, 0)])
        +- StateStoreRestore [event_time#118-T10000ms], state info [ checkpoint = <unknown>, runId = a870e6e2-b925-4104-9886-b211c0be1b73, opId = 0, ver = 0, numPartitions = 1 ], 2
            +- ObjectHashAggregate(keys=[event_time#118-T10000ms], functions=[merge_collect_list(batch#141, 0, 0), merge_collect_list(id#129, 0, 0)])
                +- Exchange hashpartitioning(event_time#118-T10000ms, 1)
                    +- ObjectHashAggregate(keys=[event_time#118-T10000ms], functions=[partial_collect_list(batch#141, 0, 0), partial_collect_list(id#129, 0, 0)])
                        +- EventTimeWatermark event_time#118: timestamp, interval 10 seconds
                            +- *(1) Project [cast(from_unixtime(cast(split(cast(value#8 as string), ,)[0] as bigint), yyyy-MM-dd HH:mm:ss, Some(Europe/Warsaw)) as timestamp) AS event_time#118, split(cast(value#8 as string), ,)[1] AS id#129, cast(split(cast(value#8 as string), ,)[2] as int) AS batch#141]
                                +- StreamingRelation kafka, [key#7, value#8, topic#9, partition#10, offset#11L, timestamp#12, timestampType#13]

val queryName = "ids-kafka"

```

```

val checkpointLocation = s"/tmp/checkpoint-$queryName"

// Delete the checkpoint location from previous executions
import java.nio.file.{Files, FileSystems}
import java.util.Comparator
import scala.collection.JavaConverters._
val path = FileSystems.getDefault.getPath(checkpointLocation)
if (Files.exists(path)) {
  Files.walk(path)
    .sorted(Comparator.reverseOrder())
    .iterator
    .asScala
    .foreach(p => p.toFile.delete)
}

// The following make for an easier demo
// Kafka cluster is supposed to be up at this point
// Make sure that a Kafka topic is available, e.g. topic-00
// Use ./bin/kafka-console-producer.sh --broker-list :9092 --topic topic-00
// And send a record, e.g. 1,1,1

// Define the output mode
// and start the query
import scala.concurrent.duration._
import org.apache.spark.sql.streaming.OutputMode.Append
import org.apache.spark.sql.streaming.Trigger
val streamingQuery = ids
  .writeStream
  .format("console")
  .option("truncate", false)
  .option("checkpointLocation", checkpointLocation)
  .queryName(queryName)
  .outputMode(Append)
  .start

val lastProgress = streamingQuery.lastProgress
scala> :type lastProgress
org.apache.spark.sql.streaming.StreamingQueryProgress

assert(lastProgress.stateOperators.length == 1, "There should be one stateful operator"
)

scala> println(lastProgress.stateOperators.head.prettyJson)
{
  "numRowsTotal" : 1,
  "numRowsUpdated" : 0,
  "memoryUsedBytes" : 742,
  "customMetrics" : {
    "loadedMapCacheHitCount" : 1,
    "loadedMapCacheMissCount" : 1,
    "stateOnCurrentVersionSizeBytes" : 374
  }
}

```

```
assert(lastProgress.sources.length == 1, "There should be one streaming source only")
scala> println(lastProgress.sources.head.prettyJson)
{
  "description" : "KafkaV2[SubscribePattern[topic-\\d{2}]]",
  "startOffset" : {
    "topic-00" : {
      "0" : 1
    }
  },
  "endOffset" : {
    "topic-00" : {
      "0" : 1
    }
  },
  "numInputRows" : 0,
  "inputRowsPerSecond" : 0.0,
  "processedRowsPerSecond" : 0.0
}

// Eventually...
streamingQuery.stop()
```

## Demo: groupByKey Streaming Aggregation in Update Mode

The example shows `Dataset.groupByKey` streaming operator to count rows in `Update` output mode.

In other words, it is an example of using `Dataset.groupByKey` with `count` aggregation function to count customer orders (`t`) per zip code (`k`).

Complete Spark Structured Streaming Application

```

package pl.japila.spark.examples

import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.streaming.{OutputMode, Trigger}

object GroupByKeyStreamingApp extends App {

    val inputTopic = "GroupByKeyApp-input"
    val appName = this.getClass.getSimpleName.replace("$", "")

    val spark = SparkSession.builder
        .master("local[*]")
        .appName(appName)
        .getOrCreate
    import spark.implicits._

    case class Order(id: Long, zipCode: String)

    // Input (source node)
    val orders = spark
        .readStream
        .format("kafka")
        .option("startingOffsets", "latest")
        .option("subscribe", inputTopic)
        .option("kafka.bootstrap.servers", ":9092")
        .load
        .select($"offset" as "id", $"value" as "zipCode") // FIXME Use csv, json, avro
        .as[Order]

    // Processing logic
    // groupByKey + count
    val byZipCode = (o: Order) => o.zipCode
    val ordersByZipCode = orders.groupByKey(byZipCode)

    import org.apache.spark.sql.functions.count
    val typedCountCol = (count("zipCode") as "count").as[String]
    val counts = ordersByZipCode
        .agg(typedCountCol)
        .select($"value" as "zip_code", $"count")

    // Output (sink node)
    import scala.concurrent.duration._
    counts
        .writeStream
        .format("console")
        .outputMode(OutputMode.Update) // FIXME Use Complete
        .queryName(appName)
        .trigger(Trigger.ProcessingTime(5.seconds))
        .start
        .awaitTermination()
}

```

## Credits

- The example with customer orders and postal codes is borrowed from Apache Beam's [Using GroupByKey Programming Guide](#).

# Demo: StateStoreSaveExec with Complete Output Mode

The following example code shows the behaviour of [StateStoreSaveExec](#) in Complete output mode.

```
// START: Only for easier debugging
// The state is then only for one partition
// which should make monitoring it easier
import org.apache.spark.sql.internal.SQLConf.SHUFFLE_PARTITIONS
spark.sessionState.conf.setConf(SHUFFLE_PARTITIONS, 1)
scala> spark.sessionState.conf.numShufflePartitions
res1: Int = 1
// END: Only for easier debugging

// Read datasets from a Kafka topic
// ./bin/spark-shell --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.3.0-SNAPS
HOT
// Streaming aggregation using groupBy operator is required to have StateStoreSaveExec
operator
val valuesPerGroup = spark.
  readStream.
  format("kafka").
  option("subscribe", "topic1").
  option("kafka.bootstrap.servers", "localhost:9092").
  load.
  withColumn("tokens", split('value, ",")).
  withColumn("group", 'tokens(0)).
  withColumn("value", 'tokens(1) cast "int").
  select("group", "value").
  groupBy($"group").
  agg(collect_list("value") as "values").
  orderBy($"group".asc)

// valuesPerGroup is a streaming Dataset with just one source
// so it knows nothing about output mode or watermark yet
// That's why StatefulOperatorStateInfo is generic
// and no batch-specific values are printed out
// That will be available after the first streaming batch
// Use sq.explain to know the runtime-specific values
scala> valuesPerGroup.explain
== Physical Plan ==
*Sort [group#25 ASC NULLS FIRST], true, 0
+- Exchange rangepartitioning(group#25 ASC NULLS FIRST, 1)
   +- ObjectHashAggregate(keys=[group#25], functions=[collect_list(value#36, 0, 0)])
      +- Exchange hashpartitioning(group#25, 1)
         +- StateStoreSave [group#25], StatefulOperatorStateInfo(<unknown>, 899f0fd1-b2
02-45cd-9ebd-09101ca90fa8, 0, 0), Append, 0
```

```

        +- ObjectHashAggregate(keys=[group#25], functions=[merge_collect_list(value#36, 0, 0)])
        +- Exchange hashpartitioning(group#25, 1)
        +- StateStoreRestore [group#25], StatefulOperatorStateInfo(<unknown>, 899f0fd1-b202-45cd-9ebd-09101ca90fa8,0,0)
        +- ObjectHashAggregate(keys=[group#25], functions=[merge_collect_list(value#36, 0, 0)])
        +- Exchange hashpartitioning(group#25, 1)
        +- ObjectHashAggregate(keys=[group#25], functions=[partial_collect_list(value#36, 0, 0)])
        +- *Project [split(cast(value#1 as string), ,)[0] AS group#25, cast(split(cast(value#1 as string), ,)[1] as int) AS value#36]
        +- StreamingRelation kafka, [key#0, value#1, topic#2, partition#3, offset#4L, timestamp#5, timestampType#6]

// Start the query and hence StateStoreSaveExec
// Use Complete output mode
import scala.concurrent.duration._
import org.apache.spark.sql.streaming.{OutputMode, Trigger}
val sq = valuesPerGroup.
  writeStream.
  format("console").
  option("truncate", false).
  trigger(Trigger.ProcessingTime(10.seconds)).
  outputMode(OutputMode.Complete).
  start

-----
Batch: 0
-----
+---+---+
|group|values|
+---+---+
+---+---+


// there's only 1 stateful operator and hence 0 for the index in stateOperators
scala> println(sq.lastProgress.stateOperators(0).prettyJson)
{
  "numRowsTotal" : 0,
  "numRowsUpdated" : 0,
  "memoryUsedBytes" : 60
}

// publish 1 new key-value pair in a single streaming batch
// 0,1

-----
Batch: 1
-----
+---+---+
|group|values|
+---+---+
|0 |[1] |

```

## StateStoreSaveExec with Complete Output Mode

```
+-----+-----+  
  
// it's Complete output mode so numRowsTotal is the number of keys in the state store  
// no keys were available earlier (it's just started!) and so numRowsUpdated is 0  
scala> println(sq.lastProgress.stateOperators(0).prettyJson)  
{  
  "numRowsTotal" : 1,  
  "numRowsUpdated" : 0,  
  "memoryUsedBytes" : 324  
}  
  
// publish new key and old key in a single streaming batch  
// new keys  
// 1,1  
// updates to already-stored keys  
// 0,2  
  
-----  
Batch: 2  
-----  
+-----+-----+  
|group|values|  
+-----+-----+  
| 0    |[2, 1]|  
| 1    |[1]     |  
+-----+-----+  
  
// it's Complete output mode so numRowsTotal is the number of keys in the state store  
// no keys were available earlier and so numRowsUpdated is...0?!  
// Think it's a BUG as it should've been 1 (for the row 0,2)  
// 8/30 Sent out a question to the Spark user mailing list  
scala> println(sq.lastProgress.stateOperators(0).prettyJson)  
{  
  "numRowsTotal" : 2,  
  "numRowsUpdated" : 0,  
  "memoryUsedBytes" : 572  
}  
  
// In the end...  
sq.stop
```

## Demo: StateStoreSaveExec with Update Output Mode

Caution	FIXME Example of Update with StateStoreSaveExec (and optional watermark)
---------	--

# Demo: Developing Custom Streaming Sink (and Monitoring SQL Queries in web UI)

The demo shows the steps to develop a custom [streaming sink](#) and use it to monitor whether and what SQL queries are executed at runtime (using web UI's SQL tab).

Note	<p>The main motivation was to answer the question <a href="#">Why does a single structured query run multiple SQL queries per batch?</a> that happened to have turned out fairly surprising.</p> <p>You're very welcome to upvote the question and answers at your earliest convenience. Thanks!</p>
------	--

The steps are as follows:

1. [Creating Custom Sink — DemoSink](#)
2. [Creating StreamSinkProvider — DemoSinkProvider](#)
3. [Optional Sink Registration using META-INF/services](#)
4. [build.sbt Definition](#)
5. [Packaging DemoSink](#)
6. [Using DemoSink in Streaming Query](#)
7. [Monitoring SQL Queries using web UI's SQL Tab](#)

Findings (aka *surprises*):

1. Custom sinks require that you define a checkpoint location using `checkpointLocation` option (or `spark.sql.streaming.checkpointLocation` Spark property). Remove the checkpoint directory (or use a different one every start of a streaming query) to have consistent results.

## Creating Custom Sink — DemoSink

A streaming sink follows the [Sink contract](#) and a sample implementation could look as follows.

```

package pl.japila.spark.sql.streaming

case class DemoSink(
    sqlContext: SQLContext,
    parameters: Map[String, String],
    partitionColumns: Seq[String],
    outputMode: OutputMode) extends Sink {

    override def addBatch(batchId: Long, data: DataFrame): Unit = {
        println(s"addBatch($batchId)")
        data.explain()
        // Why so many lines just to show the input DataFrame?
        data.sparkSession.createDataFrame(
            data.sparkSession.sparkContext.parallelize(data.collect()), data.schema)
            .show(10)
    }
}

```

Save the file under `src/main/scala` in your project.

## Creating StreamSinkProvider — DemoSinkProvider

```

package pl.japila.spark.sql.streaming

class DemoSinkProvider extends StreamSinkProvider
    with DataSourceRegister {

    override def createSink(
        sqlContext: SQLContext,
        parameters: Map[String, String],
        partitionColumns: Seq[String],
        outputMode: OutputMode): Sink = {
        DemoSink(sqlContext, parameters, partitionColumns, outputMode)
    }

    override def shortName(): String = "demo"
}

```

Save the file under `src/main/scala` in your project.

## Optional Sink Registration using META-INF/services

The step is optional, but greatly improve the experience when using the custom sink so you can use it by its name (rather than a fully-qualified class name or using a special class name for the sink provider).

Create `org.apache.spark.sql.sources.DataSourceRegister` in `META-INF/services` directory with the following content.

```
pl.japila.spark.sql.streaming.DemoSinkProvider
```

Save the file under `src/main/resources` in your project.

## build.sbt Definition

If you use my beloved build tool [sbt](#) to manage the project, use the following `build.sbt`.

```
organization := "pl.japila.spark"
name := "spark-structured-streaming-demo-sink"
version := "0.1"

scalaVersion := "2.11.11"

libraryDependencies += "org.apache.spark" %% "spark-sql" % "2.2.0"
```

## Packaging DemoSink

The step depends on what build tool you use to manage the project. Use whatever command you use to create a jar file with the above classes compiled and bundled together.

```
$ sbt package
[info] Loading settings from plugins.sbt ...
[info] Loading project definition from /Users/jacek/dev/sandbox/spark-structured-streaming-demo-sink/project
[info] Loading settings from build.sbt ...
[info] Set current project to spark-structured-streaming-demo-sink (in build file:/Users/jacek/dev/sandbox/spark-structured-streaming-demo-sink/)
[info] Compiling 1 Scala source to /Users/jacek/dev/sandbox/spark-structured-streaming-demo-sink/target/scala-2.11/classes ...
[info] Done compiling.
[info] Packaging /Users/jacek/dev/sandbox/spark-structured-streaming-demo-sink/target/scala-2.11/spark-structured-streaming-demo-sink_2.11-0.1.jar ...
[info] Done packaging.
[success] Total time: 5 s, completed Sep 12, 2017 9:34:19 AM
```

The jar with the sink is `/Users/jacek/dev/sandbox/spark-structured-streaming-demo-sink/target/scala-2.11/spark-structured-streaming-demo-sink_2.11-0.1.jar`.

## Using DemoSink in Streaming Query

The following code reads data from the `rate` source and simply outputs the result to our custom `DemoSink`.

```
// Make sure the DemoSink jar is available
$ ls /Users/jacek/dev/sandbox/spark-structured-streaming-demo-sink/target/scala-2.11/spark-structured-streaming-demo-sink_2.11-0.1.jar
/Users/jacek/dev/sandbox/spark-structured-streaming-demo-sink/target/scala-2.11/spark-structured-streaming-demo-sink_2.11-0.1.jar

// "Install" the DemoSink using --jars command-line option
$ ./bin/spark-shell --jars /Users/jacek/dev/sandbox/spark-structured-streaming-custom-sink/target/scala-2.11/spark-structured-streaming-custom-sink_2.11-0.1.jar

scala> spark.version
res0: String = 2.3.0-SNAPSHOT

import org.apache.spark.sql.streaming._
import scala.concurrent.duration._
val sq = spark.
  readStream.
  format("rate").
  load.
  writeStream.
  format("demo").
  option("checkpointLocation", "/tmp/demo-checkpoint").
  trigger(Trigger.ProcessingTime(10.seconds)).
  start

// In the end...
scala> sq.stop
17/09/12 09:59:28 INFO StreamExecution: Query [id = 03cd78e3-94e2-439c-9c12-cfed0c9968
12, runId = 6938af91-9806-4404-965a-5ae7525d5d3f] was stopped
```

## Monitoring SQL Queries using web UI's SQL Tab

Open <http://localhost:4040/SQL/>.

You should find that every trigger (aka *batch*) results in 3 SQL queries. Why?

ID	Description	Submitted	Duration	Job IDs
80	start at <console>:43	+details 2017/09/12 09:55:30	20 ms	103 104 105
79	start at <console>:43	+details 2017/09/12 09:55:30	9 ms	102
78	start at <console>:43	+details 2017/09/12 09:55:30	36 ms	
77	start at <console>:43	+details 2017/09/12 09:55:20	32 ms	99 100 101
76	start at <console>:43	+details 2017/09/12 09:55:20	9 ms	98
75	start at <console>:43	+details 2017/09/12 09:55:20	49 ms	
74	start at <console>:43	+details 2017/09/12 09:55:10	21 ms	95 96 97
73	start at <console>:43	+details 2017/09/12 09:55:10	8 ms	94
72	start at <console>:43	+details 2017/09/12 09:55:10	39 ms	
71	start at <console>:43	+details 2017/09/12 09:55:00	19 ms	91 92 93
70	start at <console>:43	+details 2017/09/12 09:55:00	9 ms	90
69	start at <console>:43	+details 2017/09/12 09:55:00	37 ms	

Figure 1. web UI's SQL Tab and Completed Queries (3 Queries per Batch)

The answer lies in what sources and sink a streaming query uses (and differs per streaming query).

In our case, `DemoSink` collects the rows from the input `DataFrame` and shows it afterwards. That gives 2 SQL queries (as you can see after executing the following batch queries).

```
// batch non-streaming query
val data = (0 to 3).toDF("id")

// That gives one SQL query
data.collect

// That gives one SQL query, too
data.show
```

The remaining query (which is the first among the queries) is executed when you load the data.

That can be observed easily when you change `DemoSink` to not "touch" the input `data` (in `addBatch`) in any way.

```
override def addBatch(batchId: Long, data: DataFrame): Unit = {
    println(s"addBatch($batchId)")
}
```

Re-run the streaming query (using the new `DemoSink`) and use web UI's SQL tab to see the queries. You should have just one query per batch (and no Spark jobs given nothing is really done in the sink's `addBatch`).

The screenshot shows the Spark 2.3.0-SNAPSHOT web UI interface. At the top, there is a navigation bar with tabs: Jobs, Stages, Storage, Environment, Executors, and SQL. The SQL tab is currently selected, indicated by a grey background. To the right of the navigation bar, it says "Spark shell application UI". Below the navigation bar, the title "SQL" is displayed in bold. Underneath "SQL", the heading "Completed Queries" is shown. A table follows, listing two completed queries:

ID	Description	Submitted	Duration	Job IDs
1	start at <console>:37	+details 2017/09/12 10:26:40	0 ms	
0	start at <console>:37	+details 2017/09/12 10:26:38	0 ms	

Figure 2. web UI's SQL Tab and Completed Queries (1 Query per Batch)

# Demo: current\_timestamp Function For Processing Time in Streaming Queries

The demo shows what happens when you use `current_timestamp` function in your structured queries.

Note

The main motivation was to answer the question [How to achieve ingestion time?](#) in Spark Structured Streaming.

You're very welcome to upvote the question and answers at your earliest convenience. Thanks!

Quoting the [Apache Flink documentation](#):

**Event time** is the time that each individual event occurred on its producing device. This time is typically embedded within the records before they enter Flink and that event timestamp can be extracted from the record.

That is exactly how event time is considered in `withWatermark` operator which you use to describe what column to use for event time. The column could be part of the input dataset or...generated.

And that is the moment where my confusion starts.

In order to generate the event time column for `withWatermark` operator you could use `current_timestamp` or `current_date` standard functions.

```
// rate format gives event time
// but let's generate a brand new column with ours
// for demo purposes
val values = spark.
  readStream.
  format("rate").
  load.
  withColumn("current_timestamp", current_timestamp)
scala> values.printSchema
root
|-- timestamp: timestamp (nullable = true)
|-- value: long (nullable = true)
|-- current_timestamp: timestamp (nullable = false)
```

Both are special for Spark Structured Streaming as `streamExecution` replaces their underlying Catalyst expressions, `CurrentTimestamp` and `CurrentDate` respectively, with `CurrentBatchTimestamp` expression and the time of the current batch.

```

import org.apache.spark.sql.streaming.Trigger
import scala.concurrent.duration._
val sq = values.
  writeStream.
  format("console").
  option("truncate", false).
  trigger(Trigger.ProcessingTime(10.seconds)).
  start

// note the value of current_timestamp
// that corresponds to the batch time

-----
Batch: 1
-----
+-----+-----+
|timestamp          |value|current_timestamp |
+-----+-----+
|2017-09-18 10:53:31.523|0    |2017-09-18 10:53:40|
|2017-09-18 10:53:32.523|1    |2017-09-18 10:53:40|
|2017-09-18 10:53:33.523|2    |2017-09-18 10:53:40|
|2017-09-18 10:53:34.523|3    |2017-09-18 10:53:40|
|2017-09-18 10:53:35.523|4    |2017-09-18 10:53:40|
|2017-09-18 10:53:36.523|5    |2017-09-18 10:53:40|
|2017-09-18 10:53:37.523|6    |2017-09-18 10:53:40|
|2017-09-18 10:53:38.523|7    |2017-09-18 10:53:40|
+-----+-----+

// Use web UI's SQL tab for the batch (Submitted column)
// or sq.recentProgress
scala> println(sq.recentProgress(1).timestamp)
2017-09-18T08:53:40.000Z

// Note current_batch_timestamp

scala> sq.explain(extended = true)
== Parsed Logical Plan ==
'Project [timestamp#2137, value#2138L, current_batch_timestamp(1505725650005, TimestampType, None) AS current_timestamp#50]
+- LogicalRDD [timestamp#2137, value#2138L], true

== Analyzed Logical Plan ==
timestamp: timestamp, value: bigint, current_timestamp: timestamp
Project [timestamp#2137, value#2138L, current_batch_timestamp(1505725650005, TimestampType, Some(Europe/Berlin)) AS current_timestamp#50]
+- LogicalRDD [timestamp#2137, value#2138L], true

== Optimized Logical Plan ==
Project [timestamp#2137, value#2138L, 1505725650005000 AS current_timestamp#50]
+- LogicalRDD [timestamp#2137, value#2138L], true

== Physical Plan ==

```

```
*Project [timestamp#2137, value#2138L, 1505725650005000 AS current_timestamp#50]
+- Scan ExistingRDD[timestamp#2137,value#2138L]
```

That *seems* to be closer to processing time than ingestion time given the definition from the [Apache Flink documentation](#):

**Processing time** refers to the system time of the machine that is executing the respective operation.

**Ingestion time** is the time that events enter Flink.

What do you think?

# Demo: Using StreamingQueryManager for Query Termination Management

The demo shows how to use [StreamingQueryManager](#) (and specifically [awaitAnyTermination](#) and [resetTerminated](#)) for query termination management.

`demo-StreamingQueryManager.scala`

```
// Save the code as demo-StreamingQueryManager.scala
// Start it using spark-shell
// $ ./bin/spark-shell -i demo-StreamingQueryManager.scala

// Register a StreamingQueryListener to receive notifications about state changes of streaming queries
import org.apache.spark.sql.streaming.StreamingQueryListener
val myQueryListener = new StreamingQueryListener {
    import org.apache.spark.sql.streaming.StreamingQueryListener._
    def onQueryTerminated(event: QueryTerminatedEvent): Unit = {
        println(s"Query ${event.id} terminated")
    }

    def onQueryStarted(event: QueryStartedEvent): Unit = {}
    def onQueryProgress(event: QueryProgressEvent): Unit = {}
}
spark.streams.addListener(myQueryListener)

import org.apache.spark.sql.streaming._
import scala.concurrent.duration._

// Start streaming queries

// Start the first query
val q4s = spark.readStream.
    format("rate").
    load.
    writeStream.
    format("console").
    trigger(Trigger.ProcessingTime(4.seconds)).
    option("truncate", false).
    start

// Start another query that is slightly slower
val q10s = spark.readStream.
    format("rate").
    load.
    writeStream.
    format("console").
    trigger(Trigger.ProcessingTime(10.seconds)).
    option("truncate", false).
```

```

start

// Both queries run concurrently
// You should see different outputs in the console
// q4s prints out 4 rows every batch and twice as often as q10s
// q10s prints out 10 rows every batch

/*
-----
Batch: 7
-----
+-----+-----+
|timestamp          |value|
+-----+-----+
|2017-10-27 13:44:07.462|21   |
|2017-10-27 13:44:08.462|22   |
|2017-10-27 13:44:09.462|23   |
|2017-10-27 13:44:10.462|24   |
+-----+-----+

-----
Batch: 8
-----
+-----+-----+
|timestamp          |value|
+-----+-----+
|2017-10-27 13:44:11.462|25   |
|2017-10-27 13:44:12.462|26   |
|2017-10-27 13:44:13.462|27   |
|2017-10-27 13:44:14.462|28   |
+-----+-----+

-----
Batch: 2
-----
+-----+-----+
|timestamp          |value|
+-----+-----+
|2017-10-27 13:44:09.847|6    |
|2017-10-27 13:44:10.847|7    |
|2017-10-27 13:44:11.847|8    |
|2017-10-27 13:44:12.847|9    |
|2017-10-27 13:44:13.847|10   |
|2017-10-27 13:44:14.847|11   |
|2017-10-27 13:44:15.847|12   |
|2017-10-27 13:44:16.847|13   |
|2017-10-27 13:44:17.847|14   |
|2017-10-27 13:44:18.847|15   |
+-----+-----+
*/
// Stop q4s on a separate thread
// as we're about to block the current thread awaiting query termination

```

```

import java.util.concurrent.Executors
import java.util.concurrent.TimeUnit.SECONDS
def queryTerminator(query: StreamingQuery) = new Runnable {
  def run = {
    println(s"Stopping streaming query: ${query.id}")
    query.stop
  }
}
import java.util.concurrent.TimeUnit.SECONDS
// Stop the first query after 10 seconds
Executors.newSingleThreadScheduledExecutor.
  scheduleWithFixedDelay(queryTerminator(q4s), 10, 60 * 5, SECONDS)
// Stop the other query after 20 seconds
Executors.newSingleThreadScheduledExecutor.
  scheduleWithFixedDelay(queryTerminator(q10s), 20, 60 * 5, SECONDS)

// Use StreamingQueryManager to wait for any query termination (either q1 or q2)
// the current thread will block indefinitely until either streaming query has finished

spark.streams.awaitAnyTermination

// You are here only after either streaming query has finished
// Executing spark.streams.awaitAnyTermination again would return immediately

// You should have received the QueryTerminatedEvent for the query termination

// reset the last terminated streaming query
spark.streams.resetTerminated

// You know at least one query has terminated

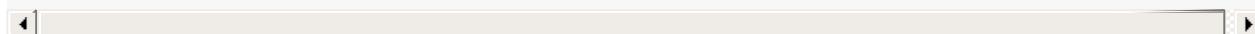
// Wait for the other query to terminate
spark.streams.awaitAnyTermination

assert(spark.streams.active.isEmpty)

println("The demo went all fine. Exiting...")

// leave spark-shell
System.exit(0)

```



# Streaming Aggregation

In Spark Structured Streaming, a **streaming aggregation** is a streaming query that was described (*build*) using the following [high-level streaming operators](#):

- `Dataset.groupBy`, `Dataset.rollup`, `Dataset.cube` (that simply create a `RelationalGroupedDataset` )
- `Dataset.groupByKey` (that simply creates a `KeyValueGroupedDataset` )
- SQL's `GROUP BY` clause (including `WITH CUBE` and `WITH ROLLUP` )

Streaming aggregation belongs to the category of [Stateful Stream Processing](#).

## IncrementalExecution — QueryExecution of Streaming Queries

Under the covers, the high-level operators create a logical query plan with one or more `Aggregate` logical operators.

Tip	Read up on <a href="#">Aggregate</a> logical operator in <a href="#">The Internals of Spark SQL</a> book.
-----	---

In Spark Structured Streaming `IncrementalExecution` is responsible for planning streaming queries for execution.

At [query planning](#), `IncrementalExecution` uses the [StatefulAggregationStrategy](#) execution planning strategy for planning streaming aggregations ( `Aggregate` unary logical operators) as pairs of [StateStoreRestoreExec](#) and [StateStoreSaveExec](#) physical operators.

```
// input data from a data source
// it's rate data source
// but that does not really matter
// We need a streaming Dataset
val input = spark
  .readStream
  .format("rate")
  .load

// Streaming aggregation with groupBy
val counts = input
  .groupBy($"value" % 2)
  .count

counts.explain(extended = true)
/** 
 == Parsed Logical Plan ==

```

```
'Aggregate [('value % 2)], [('value % 2) AS (value % 2)#23, count(1) AS count#22L]
+- StreamingRelationV2 org.apache.spark.sql.execution.streaming.sources.RateStreamProv
ider@7879348, rate, [timestamp#15, value#16L]

== Analyzed Logical Plan ==
(value % 2): bigint, count: bigint
Aggregate [(value#16L % cast(2 as bigint))], [(value#16L % cast(2 as bigint)) AS (valu
e % 2)#23L, count(1) AS count#22L]
+- StreamingRelationV2 org.apache.spark.sql.execution.streaming.sources.RateStreamProv
ider@7879348, rate, [timestamp#15, value#16L]

== Optimized Logical Plan ==
Aggregate [(value#16L % 2)], [(value#16L % 2) AS (value % 2)#23L, count(1) AS count#22
L]
+- Project [value#16L]
  +- StreamingRelationV2 org.apache.spark.sql.execution.streaming.sources.RateStreamP
rovider@7879348, rate, [timestamp#15, value#16L]

== Physical Plan ==
*(4) HashAggregate(keys=[(value#16L % 2)#27L], functions=[count(1)], output=[(value %
2)#23L, count#22L])
+- StateStoreSave [(value#16L % 2)#27L], state info [ checkpoint = <unknown>, runId =
8c0ae2be-5eaa-4038-bc29-a176abfaf885, opId = 0, ver = 0, numPartitions = 200], Append,
  0, 2
  +- *(3) HashAggregate(keys=[(value#16L % 2)#27L], functions=[merge_count(1)], outpu
t=[(value#16L % 2)#27L, count#29L])
    +- StateStoreRestore [(value#16L % 2)#27L], state info [ checkpoint = <unknown>,
      runId = 8c0ae2be-5eaa-4038-bc29-a176abfaf885, opId = 0, ver = 0, numPartitions = 200]
    , 2
    +- *(2) HashAggregate(keys=[(value#16L % 2)#27L], functions=[merge_count(1)],
      output=[(value#16L % 2)#27L, count#29L])
      +- Exchange hashpartitioning((value#16L % 2)#27L, 200)
        +- *(1) HashAggregate(keys=[(value#16L % 2) AS (value#16L % 2)#27L], fu
nctions=[partial_count(1)], output=[(value#16L % 2)#27L, count#29L])
          +- *(1) Project [value#16L]
            +- StreamingRelation rate, [timestamp#15, value#16L]
*/

```

## Demos

Use the following demos to learn more:

- [Streaming Watermark with Aggregation in Append Output Mode](#)
- [Streaming Query for Running Counts \(Socket Source and Complete Output Mode\)](#)
- [Streaming Aggregation with Kafka Data Source](#)
- [groupByKey Streaming Aggregation in Update Mode](#)



# StateStoreRDD — RDD for Updating State (in StateStores Across Spark Cluster)

`StateStoreRDD` is an `RDD` for executing `storeUpdateFunction` with `StateStore` (and data from partitions of the `data RDD`).

`StateStoreRDD` is created for the following stateful physical operators (using `StateStoreOps.mapPartitionsWithStateStore`):

- `FlatMapGroupsWithStateExec`
- `StateStoreRestoreExec`
- `StateStoreSaveExec`
- `StreamingDeduplicateExec`
- `StreamingGlobalLimitExec`

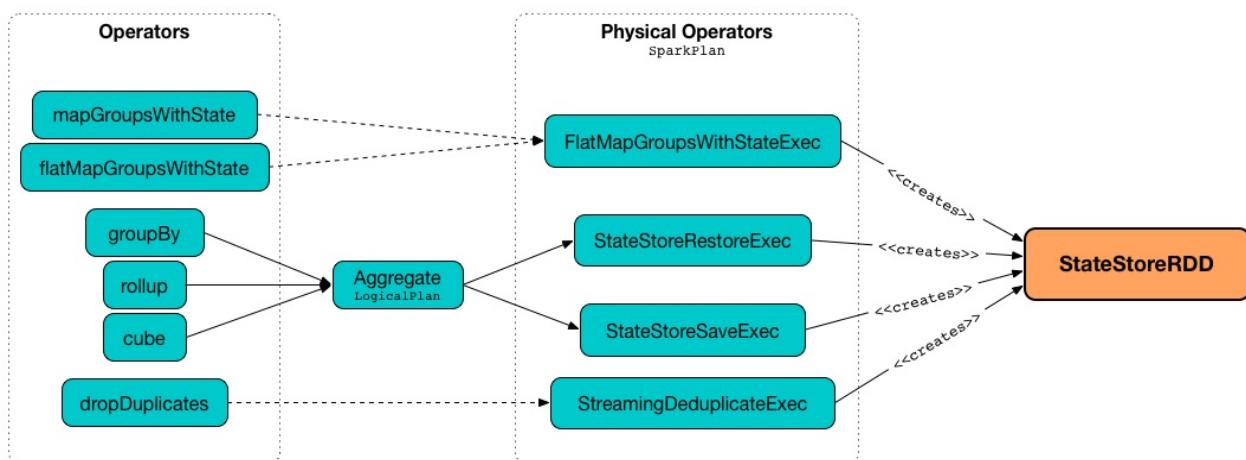


Figure 1. StateStoreRDD, Physical and Logical Plans, and operators

`StateStoreRDD` uses `StateStoreCoordinator` for the preferred locations of a partition for job scheduling.

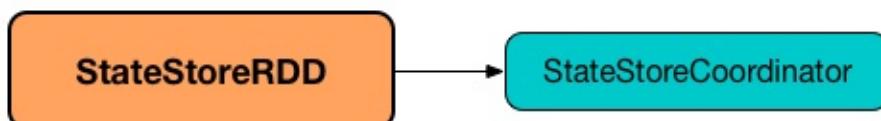


Figure 2. StateStoreRDD and StateStoreCoordinator

`getPartitions` is exactly the partitions of the `data RDD`.

## Computing Partition — `compute` Method

```
compute(
    partition: Partition,
    ctxt: TaskContext): Iterator[U]
```

**Note** `compute` is part of the RDD Contract to compute a given partition.

`compute` computes `dataRDD` passing the result on to `storeUpdateFunction` (with a configured `StateStore`).

Internally, (and similarly to `getPreferredLocations`) `compute` creates a `StateStoreProviderId` with `statestoreId` (using `checkpointLocation`, `operatorId` and the index of the input `partition`) and `queryRunId`.

`compute` then requests `statestore` for the store for the `StateStoreProviderId`.

In the end, `compute` computes `dataRDD` (using the input `partition` and `ctxt`) followed by executing `storeUpdateFunction` (with the store and the result).

## Placement Preferences of Partition (Preferred Locations)

### — `getPreferredLocations` Method

```
getPreferredLocations(partition: Partition): Seq[String]
```

**Note** `getPreferredLocations` is a part of the RDD Contract to specify placement preferences (aka *preferred task locations*), i.e. where tasks should be executed to be as close to the data as possible.

`getPreferredLocations` creates a `StateStoreProviderId` with `statestoreId` (using `checkpointLocation`, `operatorId` and the index of the input `partition`) and `queryRunId`.

**Note** `checkpointLocation` and `operatorId` are shared across different partitions and so the only difference in `StateStoreProviderIds` is the partition index.

In the end, `getPreferredLocations` requests `StateStoreCoordinatorRef` for the location of the `state store` for the `StateStoreProviderId`.

**Note** `StateStoreCoordinator` coordinates instances of `statestores` across Spark executors in the cluster, and tracks their locations for job scheduling.

## Creating StateStoreRDD Instance

`StateStoreRDD` takes the following to be created:

- Data RDD (`RDD[T]` to update the aggregates in a state store)

- Store update function ( `(stateStore, Iterator[T]) → Iterator[U]` where `T` is the type of rows in the [data RDD](#))
- Checkpoint directory
- Run ID of the streaming query
- Operator ID
- Version of the store
- **Key schema** - schema of the keys
- **Value schema** - schema of the values
- Index
- `SessionState`
- Optional [StateStoreCoordinatorRef](#)

`StateStoreRDD` initializes the [internal properties](#).

## Internal Properties

Name	Description
<code>hadoopConfBroadcast</code>	
<code>storeConf</code>	Configuration parameters (as <code>stateStoreConf</code> ) using the current <code>SQLConf</code> (from <code>SessionState</code> )

# StateStoreOps — Extension Methods for Creating StateStoreRDD

`StateStoreOps` is a **Scala implicit class** of a data RDD (of type `RDD[T]`) to [create a `StateStoreRDD`](#) for the following physical operators:

- [FlatMapGroupsWithStateExec](#)
- [StateStoreRestoreExec](#)
- [StateStoreSaveExec](#)
- [StreamingDeduplicateExec](#)

Note

[Implicit Classes](#) are a language feature in Scala for **implicit conversions** with **extension methods** for existing types.

## Creating StateStoreRDD (with `storeUpdateFunction` Aborting StateStore When Task Fails)

### — `mapPartitionsWithStateStore` Method

```
mapPartitionsWithStateStore[U](
    stateInfo: StatefulOperatorStateInfo,
    keySchema: StructType,
    valueSchema: StructType,
    indexOrdinal: Option[Int],
    sessionState: SessionState,
    storeCoordinator: Option[StateStoreCoordinatorRef])(  

    storeUpdateFunction: (StateStore, Iterator[T]) => Iterator[U]): StateStoreRDD[T, U]  

// Used for testing only  

mapPartitionsWithStateStore[U](
    sqlContext: SQLContext,
    stateInfo: StatefulOperatorStateInfo,
    keySchema: StructType,
    valueSchema: StructType,
    indexOrdinal: Option[Int])(  

    storeUpdateFunction: (StateStore, Iterator[T]) => Iterator[U]): StateStoreRDD[T, U] (1)
```

1. Uses `sqlContext.streams.stateStoreCoordinator` to access `StateStoreCoordinator` Internally, `mapPartitionsWithStateStore` requests `SparkContext` to clean `storeUpdateFunction` function.

Note	<code>mapPartitionsWithStateStore</code> uses the enclosing RDD to access the current SparkContext .
------	--

Note	<b>Function Cleaning</b> is to clean a closure from unreferenced variables before it is serialized and sent to tasks. <code>SparkContext</code> reports a <code>SparkException</code> when the closure is not serializable.
------	---

`mapPartitionsWithStateStore` then creates a (wrapper) function to `abort` the `StateStore` if `state updates had not been committed` before a task finished (which is to make sure that the `StateStore` has been `committed` or `aborted` in the end to follow the contract of `StateStore` ).

Note	<code>mapPartitionsWithStateStore</code> uses <code>TaskCompletionListener</code> to be notified when a task has finished.
------	--

In the end, `mapPartitionsWithStateStore` creates a `StateStoreRDD` (with the wrapper function, `SessionState` and `StateStoreCoordinatorRef`).

Note	<code>mapPartitionsWithStateStore</code> is used when the following physical operators are executed: <ul style="list-style-type: none"> <li>• <code>FlatMapGroupsWithStateExec</code></li> <li>• <code>StateStoreRestoreExec</code></li> <li>• <code>StateStoreSaveExec</code></li> <li>• <code>StreamingDeduplicateExec</code></li> <li>• <code>StreamingGlobalLimitExec</code></li> </ul>
------	---

# StreamingAggregationStateManager Contract — State Managers for Streaming Aggregation

`StreamingAggregationStateManager` is the [abstraction](#) of [state managers](#) that act as [middlemen](#) between [state stores](#) and the physical operators used in [Streaming Aggregation](#) (e.g. [StateStoreSaveExec](#) and [StateStoreRestoreExec](#)).

Table 1. StreamingAggregationStateManager Contract

Method	Description
<code>commit</code>	<pre>commit(     store: StateStore): Long</pre> <p>Commits all updates (<i>changes</i>) to the given <a href="#">state store</a> and returns the new version</p> <p>Used exclusively when <a href="#">StateStoreSaveExec</a> physical operator is executed.</p>
<code>get</code>	<pre>get(store: StateStore, key: UnsafeRow): UnsafeRow</pre> <p>Looks up the value of the key from the <a href="#">state store</a> (the key is non-<code>null</code> )</p> <p>Used exclusively when <a href="#">StateStoreRestoreExec</a> physical operator is executed.</p>
<code>getKey</code>	<pre>getKey(row: UnsafeRow): UnsafeRow</pre> <p>Extracts the columns for the key from the input row</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <a href="#">StateStoreRestoreExec</a> physical operator is executed</li> <li>• <code>StreamingAggregationStateManagerImplV1</code> legacy state manager is requested to <a href="#">put a row to a state store</a></li> </ul>
<code>getStateValueSchema</code>	<pre>getStateValueSchema: StructType</pre> <p>Gets the schema of the values in a <a href="#">state store</a></p>

	Used when <a href="#">StateStoreRestoreExec</a> and <a href="#">StateStoreSaveExec</a> physical operators are executed
<code>iterator</code>	<pre>iterator(     store: StateStore): Iterator[UnsafeRowPair]</pre> <p>Returns all <code>UnsafeRow</code> key-value pairs in the given <a href="#">state store</a></p> <p>Used exclusively when <a href="#">StateStoreSaveExec</a> physical operator is executed.</p>
<code>keys</code>	<pre>keys(store: StateStore): Iterator[UnsafeRow]</pre> <p>Returns all the keys in the <a href="#">state store</a></p> <p>Used exclusively when physical operators with <code>WatermarkSupport</code> are requested to <a href="#">removeKeysOlderThanWatermark</a> (i.e. exclusively when <a href="#">StateStoreSaveExec</a> physical operator is executed).</p>
<code>put</code>	<pre>put(     store: StateStore,     row: UnsafeRow): Unit</pre> <p>Stores (<i>puts</i>) the given row in the given <a href="#">state store</a></p> <p>Used exclusively when <a href="#">StateStoreSaveExec</a> physical operator is executed.</p>
<code>remove</code>	<pre>remove(     store: StateStore,     key: UnsafeRow): Unit</pre> <p>Removes the key-value pair from the given <a href="#">state store</a> per key</p> <p>Used exclusively when <a href="#">StateStoreSaveExec</a> physical operator is executed (directly or indirectly as a <a href="#">WatermarkSupport</a>)</p>
<code>values</code>	<pre>values(     store: StateStore): Iterator[UnsafeRow]</pre> <p>All values in the <a href="#">state store</a></p> <p>Used exclusively when <a href="#">StateStoreSaveExec</a> physical operator is executed.</p>

`StreamingAggregationStateManager` supports two versions of state managers for streaming aggregations (per the `spark.sql.streaming.aggregation.stateFormatVersion` internal configuration property):

- 1 (for the legacy `StreamingAggregationStateManagerImplV1`)
- 2 (for the default `StreamingAggregationStateManagerImplV2`)

Note	<code>StreamingAggregationStateManagerBaseImpl</code> is the one and only known direct implementation of the <code>StreamingAggregationStateManager Contract</code> in Spark Structured Streaming.
------	--

Note	<code>StreamingAggregationStateManager</code> is a Scala <b>sealed trait</b> which means that all the <code>implementations</code> are in the same compilation unit (a single file).
------	--

## Creating `StreamingAggregationStateManager` Instance — `createStateManager` Factory Method

```
createStateManager(  
    keyExpressions: Seq[Attribute],  
    inputRowAttributes: Seq[Attribute],  
    stateFormatVersion: Int): StreamingAggregationStateManager
```

`createStateManager` creates a new `StreamingAggregationStateManager` for a given `stateFormatVersion`:

- `StreamingAggregationStateManagerImplV1` for `stateFormatVersion` being 1
- `StreamingAggregationStateManagerImplV2` for `stateFormatVersion` being 2

`createStateManager` throws a `IllegalArgumentException` for any other `stateFormatVersion`:

```
Version [stateFormatVersion] is invalid
```

Note	<code>createStateManager</code> is used when <code>StateStoreRestoreExec</code> and <code>StateStoreSaveExec</code> physical operators are created.
------	---

# StreamingAggregationStateManagerBaseImpl

## — Base State Manager for Streaming Aggregation

`StreamingAggregationStateManagerBaseImpl` is the base implementation of the `StreamingAggregationStateManager` contract for state managers for streaming aggregations that use `UnsafeProjection` to `getKey`.

`StreamingAggregationStateManagerBaseImpl` uses `UnsafeProjection` to `getKey`.

Table 1. `StreamingAggregationStateManagerBaseImpls`

<b>StreamingAggregationStateManagerBaseImpl</b>	<b>Description</b>
<code>StreamingAggregationStateManagerImplV1</code>	Legacy <code>StreamingAggregationStateManager</code> when <code>spark.sql.streaming.aggregation.state</code> configuration property is <code>1</code> )
<code>StreamingAggregationStateManagerImplV2</code>	Default <code>StreamingAggregationStateManager</code> when <code>spark.sql.streaming.aggregation.state</code> configuration property is <code>2</code> )

`StreamingAggregationStateManagerBaseImpl` takes the following to be created:

- Catalyst expressions for the keys (`Seq[Attribute]`)
- Catalyst expressions for the input rows (`Seq[Attribute]`)

Note	<code>StreamingAggregationStateManagerBaseImpl</code> is a Scala abstract class and cannot be <code>created</code> directly. It is created indirectly for the <code>concrete</code> <code>StreamingAggregationStateManagerBaseImpls</code> .
------	--

## Committing (Changes to) State Store — `commit` Method

```
commit(  
    store: StateStore): Long
```

Note	<code>commit</code> is part of the <code>StreamingAggregationStateManager Contract</code> to commit changes to a <code>state store</code> .
------	---

`commit` simply requests the `state store` to `commit` state changes.

## Removing Key From State Store — `remove` Method

```
remove(store: StateStore, key: UnsafeRow): Unit
```

Note

`remove` is part of the [StreamingAggregationStateManager Contract](#) to remove a key from a state store.

`remove` ...FIXME

## getKey Method

```
getKey(row: UnsafeRow): UnsafeRow
```

Note

`getKey` is part of the [StreamingAggregationStateManager Contract](#) to...FIXME

`getKey` ...FIXME

## Getting All Keys in State Store — `keys` Method

```
keys(store: StateStore): Iterator[UnsafeRow]
```

Note

`keys` is part of the [StreamingAggregationStateManager Contract](#) to get all keys in a state store (as an iterator).

`keys` ...FIXME

# StreamingAggregationStateManagerImplV1 — Legacy State Manager for Streaming Aggregation

`StreamingAggregationStateManagerImplV1` is the legacy state manager for streaming aggregations.

Note	The version of a state manager is controlled using <code>spark.sql.streaming.aggregation.stateFormatVersion</code> internal configuration property.
------	---

`StreamingAggregationStateManagerImplV1` is created exclusively when `StreamingAggregationStateManager` is requested for a new `StreamingAggregationStateManager`.

## Storing Row in State Store — `put` Method

```
put(store: StateStore, row: UnsafeRow): Unit
```

Note	<code>put</code> is part of the <code>StreamingAggregationStateManager Contract</code> to store a row in a state store.
------	---

`put` ...FIXME

## Creating `StreamingAggregationStateManagerImplV1` Instance

`StreamingAggregationStateManagerImplV1` takes the following when created:

- Attribute expressions for keys (`Seq[Attribute]`)
- Attribute expressions of input rows (`Seq[Attribute]`)

# StreamingAggregationStateManagerImplV2 — Default State Manager for Streaming Aggregation

`StreamingAggregationStateManagerImplV2` is the default state manager for streaming aggregations.

Note	The version of a state manager is controlled using <code>spark.sql.streaming.aggregation.stateFormatVersion</code> internal configuration property.
------	---

`StreamingAggregationStateManagerImplV2` is created exclusively when `StreamingAggregationStateManager` is requested for a new `StreamingAggregationStateManager`.

`StreamingAggregationStateManagerImplV2` (like the parent `StreamingAggregationStateManagerBaseImpl`) takes the following to be created:

- Catalyst expressions for the keys ( `Seq[Attribute]` )
- Catalyst expressions for the input rows ( `Seq[Attribute]` )

## Storing Row in State Store — put Method

```
put(store: StateStore, row: UnsafeRow): Unit
```

Note	<code>put</code> is part of the <code>StreamingAggregationStateManager Contract</code> to store a row in a state store.
------	---

`put` ...FIXME

## Getting Saved State for Non-Null Key from State Store — get Method

```
get(store: StateStore, key: UnsafeRow): UnsafeRow
```

Note	<code>get</code> is part of the <code>StreamingAggregationStateManager Contract</code> to get the saved state for a given non-null key from a given state store.
------	--

`get` requests the given `StateStore` for the current state value for the given key.

`get` returns `null` if the key could not be found in the state store. Otherwise, `get restoreOriginalRow` (for the key and the saved state).

## restoreOriginalRow Internal Method

```
restoreOriginalRow(key: UnsafeRow, value: UnsafeRow): UnsafeRow
restoreOriginalRow(rowPair: UnsafeRowPair): UnsafeRow
```

`restoreOriginalRow ...FIXME`

Note	<code>restoreOriginalRow</code> is used when <code>StreamingAggregationStateManagerImplV2</code> is requested to <a href="#">get the saved state for a given non-null key from a state store</a> , <a href="#">iterator</a> and <a href="#">values</a> .
------	--

## getStateValueSchema Method

```
getStateValueSchema: StructType
```

Note	<code>getStateValueSchema</code> is part of the <a href="#">StreamingAggregationStateManager Contract</a> to... FIXME.
------	---

`getStateValueSchema` simply requests the [valueExpressions](#) for the schema.

## iterator Method

```
iterator: iterator(store: StateStore): Iterator[UnsafeRowPair]
```

Note	<code>iterator</code> is part of the <a href="#">StreamingAggregationStateManager Contract</a> to... FIXME.
------	--

`iterator` simply requests the input [state store](#) for the [iterator](#) that is mapped to an iterator of [UnsafeRowPairs](#) with the key (of the input [UnsafeRowPair](#)) and the value as a [restored original row](#).

Note	<code>scala.collection.Iterator</code> is a data structure that allows to iterate over a sequence of elements that are usually fetched lazily (i.e. no elements are fetched from the underlying store until processed).
------	---

## values Method

```
values(store: StateStore): Iterator[UnsafeRow]
```

**Note**

`values` is part of the [StreamingAggregationStateManager Contract](#) to...FIXME.

`values` ...FIXME

## Internal Properties

Name	Description
joiner	
keyValueJoinedExpressions	
needToProjectToRestoreValue	
restoreValueProjector	
valueExpressions	
valueProjector	

# Stateful Stream Processing

**Stateful Stream Processing** is a stream processing with state (implicit or explicit).

In Spark Structured Streaming, a streaming query is stateful when it is one of the following (that makes use of [StateStores](#)):

- [Streaming Aggregation](#)
- [Arbitrary Stateful Streaming Aggregation](#)
- [Stream-Stream Join](#)
- [Streaming Deduplication](#)
- [Streaming Limit](#)

## Versioned State, StateStores and StateStoreProviders

Spark Structured Streaming uses [StateStores](#) for versioned and fault-tolerant key-value state stores.

State stores are checkpointed incrementally to avoid state loss and for increased performance.

State stores are managed by [State Store Providers](#) with [HDFSBackedStateStoreProvider](#) being the default and only known implementation. [HDFSBackedStateStoreProvider](#) uses Hadoop DFS-compliant file system for [state checkpointing and fault-tolerance](#).

State store providers manage versioned state per [stateful operator](#) (and partition it operates on).

The lifecycle of a `stateStoreProvider` begins when `stateStore` utility (on a Spark executor) is requested for the [StateStore by provider ID and version](#).

Important	<p>It is worth to notice that since <code>stateStore</code> and <code>StateStoreProvider</code> utilities are Scala objects that makes it possible that there can only be one instance of <code>stateStore</code> and <code>StateStoreProvider</code> on a single JVM. Scala objects are (sort of) singletons which means that there will be exactly one instance of each per JVM and that is exactly the JVM of a Spark executor. As long as the executor is up and running state versions are cached and no Hadoop DFS is used (except for the initial load).</p>
-----------	---

When requested for a `StateStore`, `statestore` utility is given the version of a state store to look up. The version is either the `current epoch` (in [Continuous Stream Processing](#)) or the `current batch ID` (in [Micro-Batch Stream Processing](#)).

`StateStore` utility requests `StateStoreProvider` utility to `createAndInit` that creates the `StateStoreProvider` implementation (based on `spark.sql.streaming.stateStore.providerClass` internal configuration property) and requests it to `initialize`.

The initialized `StateStoreProvider` is cached in `loadedProviders` internal lookup table (for a `StatestoreId`) for later lookups.

`StateStoreProvider` utility then requests the `StateStoreProvider` for the `state` store for a specified version. (e.g. a `HDFSBackedStateStore` in case of `HDFSBackedStateStoreProvider`).

An instance of `StateStoreProvider` is requested to `do its own maintenance` or `close` (when a corresponding `StateStore` is `inactive`) in `MaintenanceTask` daemon thread that runs periodically every `spark.sql.streaming.stateStore.maintenanceInterval` configuration property (default: `60s` ).

## IncrementalExecution — QueryExecution of Streaming Queries

Regardless of the query language ([Dataset API](#) or SQL), any structured query (incl. streaming queries) becomes a logical query plan.

In Spark Structured Streaming it is [IncrementalExecution](#) that plans streaming queries for execution.

While [planning a streaming query for execution](#) (aka *query planning*), `IncrementalExecution` uses the [state preparation rule](#). The rule fills out the following physical operators with the execution-specific configuration (with `StatefulOperatorStateInfo` being the most important for stateful stream processing):

- `FlatMapGroupsWithStateExec`
- `StateStoreRestoreExec`
- `StateStoreSaveExec` (used for [streaming aggregation](#))
- `StreamingDeduplicateExec`
- `StreamingGlobalLimitExec`
- `StreamingSymmetricHashJoinExec`

## Micro-Batch Stream Processing and Extra Non-Data Batch for StateStoreWriter Stateful Operators

In [Micro-Batch Stream Processing](#) (with [MicroBatchExecution engine](#)), `IncrementalExecution` uses `shouldRunAnotherBatch` flag that allows `StateStoreWriters` stateful physical operators to indicate whether the last batch execution requires another non-data batch.

The following table shows the `StateStoreWriters` that redefine `shouldRunAnotherBatch` flag.

Table 1. StateStoreWriters and shouldRunAnotherBatch Flag

StateStoreWriter	shouldRunAnotherBatch Flag
<code>FlatMapGroupsWithStateExec</code>	Based on <a href="#">GroupStateTimeout</a>
<code>StateStoreSaveExec</code>	Based on <a href="#">OutputMode</a> and event-time watermark
<code>StreamingDeduplicateExec</code>	Based on event-time watermark
<code>StreamingSymmetricHashJoinExec</code>	Based on event-time watermark

## StateStoreRDD

Right after [query planning](#), a stateful streaming query (a single micro-batch actually) becomes an RDD with one or more `StateStoreRDDs`.

You can find the `StateStoreRDDs` of a streaming query in the RDD lineage.

```
scala> :type streamingQuery
org.apache.spark.sql.streaming.StreamingQuery

scala> streamingQuery.explain
== Physical Plan ==
*(4) HashAggregate(keys=[window#13-T0ms, value#3L], functions=[count(1)])
+- StateStoreSave [window#13-T0ms, value#3L], state info [ checkpoint = file:/tmp/checkpoint-counts/state, runId = 1dec2d81-f2d0-45b9-8f16-39ede66e13e7, opId = 0, ver = 1,
  numPartitions = 1], Append, 10000, 2
  +- *(3) HashAggregate(keys=[window#13-T0ms, value#3L], functions=[merge_count(1)])
    +- StateStoreRestore [window#13-T0ms, value#3L], state info [ checkpoint = file:/tmp/checkpoint-counts/state, runId = 1dec2d81-f2d0-45b9-8f16-39ede66e13e7, opId = 0,
      ver = 1, numPartitions = 1], 2
    +- *(2) HashAggregate(keys=[window#13-T0ms, value#3L], functions=[merge_count(1)])
      +- Exchange hashpartitioning(window#13-T0ms, value#3L, 1)
        +- *(1) HashAggregate(keys=[window#13-T0ms, value#3L], functions=[partial_count(1)])
```

```

        +- *(1) Project [named_struct(start, precisetimestampconversion((((CASE WHEN (cast(CEIL((cast((precisetimestampconversion(time#2-T0ms, TimestampType, LongType) - 0) as double) / 5000000.0)) as double) = (cast((precisetimestampconversion(time#2-T0ms, TimestampType, LongType) - 0) as double) / 5000000.0)) THEN (CEIL((cast((precisetimestampconversion(time#2-T0ms, TimestampType, LongType) - 0) as double) / 50000.0)) + 1) ELSE CEIL((cast((precisetimestampconversion(time#2-T0ms, TimestampType, LongType) - 0) as double) / 5000000.0)) END + 0) - 1) * 5000000) + 0), LongType, TimestampType), end, precisetimestampconversion((((CASE WHEN (cast(CEIL((cast((precisetimestampconversion(time#2-T0ms, TimestampType, LongType) - 0) as double) / 5000000.0)) as double) = (cast((precisetimestampconversion(time#2-T0ms, TimestampType, LongType) - 0) as double) / 5000000.0)) THEN (CEIL((cast((precisetimestampconversion(time#2-T0ms, TimestampType, LongType) - 0) as double) / 5000000.0)) + 1) ELSE CEIL((cast((precisetimestampconversion(time#2-T0ms, TimestampType, LongType) - 0) as double) / 5000000.0)) END + 0) - 1) * 5000000) + 5000000), LongType, TimestampType)) AS window#13-T0ms, value#3L]
        +- *(1) Filter isnotnull(time#2-T0ms)
        +- EventTimeWatermark time#2: timestamp, interval
        +- LocalTableScan <empty>, [time#2, value#3L]

import org.apache.spark.sql.execution.streaming.{StreamExecution, StreamingQueryWrapper}
}
val se = streamingQuery.asInstanceOf[StreamingQueryWrapper].streamingQuery

scala> :type se
org.apache.spark.sql.execution.streaming.StreamExecution

scala> :type se.lastExecution
org.apache.spark.sql.execution.streaming.IncrementalExecution

val rdd = se.lastExecution.toRdd
scala> rdd.toDebugString
res3: String =
(1) MapPartitionsRDD[39] at toRdd at <console>:40 []
| StateStoreRDD[38] at toRdd at <console>:40 [] // <-- here
| MapPartitionsRDD[37] at toRdd at <console>:40 []
| StateStoreRDD[36] at toRdd at <console>:40 [] // <-- here
| MapPartitionsRDD[35] at toRdd at <console>:40 []
| ShuffledRowRDD[17] at start at <pastie>:67 []
+- (1) MapPartitionsRDD[16] at start at <pastie>:67 []
| MapPartitionsRDD[15] at start at <pastie>:67 []
| MapPartitionsRDD[14] at start at <pastie>:67 []
| MapPartitionsRDD[13] at start at <pastie>:67 []
| ParallelCollectionRDD[12] at start at <pastie>:67 []

```

## StateStoreCoordinator RPC Endpoint, StateStoreRDD and Preferred Locations

Since execution of a stateful streaming query happens on Spark executors whereas planning is on the driver, Spark Structured Streaming uses RPC environment for tracking locations of the state stores in use. That makes the tasks (of a structured query) to be scheduled where the state (of a partition) is.

When planned for execution, the `StateStoreRDD` is first asked for the [preferred locations of a partition](#) (which happens on the driver) that are later used to [compute it](#) (on Spark executors).

Spark Structured Streaming uses RPC environment to keep track of [StateStores](#) (their [StateStoreProvider](#) actually) for RDD planning.

Every time [StateStoreRDD](#) is requested for the [preferred locations of a partition](#), it communicates with the [StateStoreCoordinator RPC endpoint](#) that knows the locations of the required `StateStores` (per host and executor ID).

`StateStoreRDD` uses [StateStoreProviderId](#) with [StateStoreId](#) to uniquely identify the [state store](#) to use for (*associate with*) a stateful operator and a partition.

## State Management

The state in a stateful streaming query can be implicit or explicit.

# Streaming Watermark

**Streaming Watermark** of a [stateful streaming query](#) is how long to wait for late and possibly out-of-order events until a streaming state can be considered final and not to change. Streaming watermark is used to mark events (modeled as a row in the streaming Dataset) that are older than the threshold as "too late", and not "interesting" to update partial non-final streaming state.

In Spark Structured Streaming, streaming watermark is defined using [Dataset.withWatermark](#) high-level operator.

```
withWatermark(  
    eventTime: String,  
    delayThreshold: String): Dataset[T]
```

In [Dataset.withWatermark](#) operator, `eventTime` is the name of the column to use to monitor event time whereas `delayThreshold` is a delay threshold.

**Watermark Delay** says how late and possibly out-of-order events are still acceptable and contribute to the final result of a stateful streaming query. Event-time watermark delay is used to calculate the difference between the event time of an event and the time in the past.

**Event-Time Watermark** is then a **time threshold** (*point in time*) that is the minimum acceptable time of an event (modeled as a row in the streaming Dataset) that is accepted in a stateful streaming query.

With streaming watermark, memory usage of a streaming state can be controlled as late events can easily be dropped, and old state (e.g. aggregates or join) that are never going to be updated removed. That avoids unbounded streaming state that would inevitably use up all the available memory of long-running streaming queries and end up in out of memory errors.

In [Append](#) output mode the current event-time streaming watermark is used for the following:

- Output saved state rows that became expired (**Expired events** in the demo)
- Dropping late events, i.e. don't save them to a state store or include in aggregation (**Late events** in the demo)

Streaming watermark is [required](#) for a [streaming aggregation](#) in [append](#) output mode.

## Streaming Aggregation

In [streaming aggregation](#), a streaming watermark has to be defined on one or many grouping expressions of a streaming aggregation (directly or using [window](#) standard function).

**Note**

`Dataset.withWatermark` operator has to be used before an aggregation operator (for the watermark to have an effect).

## Streaming Join

In [streaming join](#), a streaming watermark can be defined on [join keys](#) or [any of the join sides](#).

## Demos

Use the following demos to learn more:

- [Demo: Streaming Watermark with Aggregation in Append Output Mode](#)

## Internals

Under the covers, `Dataset.withWatermark` high-level operator creates a logical query plan with `EventTimeWatermark` logical operator.

`EventTimeWatermark` logical operator is planned to `EventTimeWatermarkExec` physical operator that extracts the event times (from the data being processed) and adds them to an accumulator.

Since the execution (data processing) happens on Spark executors, using the accumulator is the only *Spark-approved way* for communication between the tasks (on the executors) and the driver. Using accumulator updates the driver with the current event-time watermark.

During the query planning phase (in `MicroBatchExecution` and `ContinuousExecution`) that also happens on the driver, `IncrementalExecution` is given the current `OffsetSeqMetadata` with the current event-time watermark.

## Further Reading Or Watching

- [SPARK-18124 Observed delay based event time watermarks](#)



# Streaming Deduplication

**Streaming Deduplication** is...FIXME

# Streaming Limit

**Streaming Limit** is...FIXME

# StateStore Contract — Kay-Value Store for Streaming State Data

`StateStore` is the abstraction of key-value stores for managing state in Stateful Stream Processing (e.g. for persisting running aggregates in Streaming Aggregation).

`StateStore` supports incremental checkpointing in which only the key-value "Row" pairs that changed are committed or aborted (without touching other key-value pairs).

`StateStore` is identified with the aggregating operator id and the partition id (among other properties for identification).

Note

`HDFSBackedStateStore` is the default and only known implementation of the StateStore Contract in Spark Structured Streaming.

Table 1. StateStore Contract

Method	Description
abort	<p><code>abort(): Unit</code></p> <p>Aborts (<i>discards</i>) changes to the state store</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>StateStoreOps</code> implicit class is requested to <code>mapPartitionsWithStateStore</code> (when the state store has not been committed for a task that finishes, possibly with an error)</li> <li>• <code>StateStoreHandler</code> (of <code>SymmetricHashJoinStateManager</code>) is requested to <code>abortIfNeeded</code> (when the state store has not been committed for a task that finishes, possibly with an error)</li> </ul>
commit	<p><code>commit(): Long</code></p> <p>Commits the changes to the state store (and returns the current version)</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>FlatMapGroupsWithStateExec</code>, <code>StreamingDeduplicateExec</code> and <code>StreamingGlobalLimitExec</code> physical operators are executed (right after all rows in a partition have been processed)</li> </ul>

- `StreamingAggregationStateManagerBaseImpl` is requested to [commit \(changes to\) a state store](#) (exclusively when `StateStoreSaveExec` physical operator is executed)
- `StateStoreHandler` (of `SymmetricHashJoinStateManager`) is requested to commit changes to a state store

```
get(key: UnsafeRow): UnsafeRow
```

Looks up ([gets](#)) the value of the given non- `null` key

Used when:

- `StreamingDeduplicateExec` and `StreamingGlobalLimitExec` physical operators are executed
- `StateManagerImplBase` (of `FlatMapGroupsWithStateExecHelper`) is requested to `getState`
- `StreamingAggregationStateManagerImplV1` and `StreamingAggregationStateManagerImplV2` are requested to get the value of a non-null key
- `KeyToNumValuesStore` is requested to [get](#)
- `KeyWithIndexToValueStore` is requested to [get](#) and [getAll](#)

```
getRange(  
    start: Option[UnsafeRow],  
    end: Option[UnsafeRow]): Iterator[UnsafeRowPair]
```

Gets the key-value pairs of `UnsafeRows` for the specified range (with optional approximate `start` and `end` extents)

Used when:

`getRange`

- `WatermarkSupport` is requested to [removeKeysOlderThanWatermark](#)
- `StateManagerImplBase` is requested to `getAllState`
- `StreamingAggregationStateManagerBaseImpl` is requested to `keys`
- `KeyToNumValuesStore` and `KeyWithIndexToValueStore` are requested to `iterator`

Note

All the uses above assume the `start` and `end` as `None` that basically is `iterator`.

	<pre>hasCommitted: Boolean</pre> <p>Flag to indicate whether state changes have been committed (<code>true</code>) or not (<code>false</code>)</p>
hasCommitted	<p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>RDD</code> (via <code>stateStoreOps</code> implicit class) is requested to <code>mapPartitionsWithStateStore</code> (and a task finishes and may need to <code>abort state updates</code>)</li> <li>• <code>SymmetricHashJoinStateManager</code> is requested to <code>abortIfNeeded</code> (when a task finishes and may need to <code>abort state updates</code>)</li> </ul>
	<pre>id: StateStoreId</pre> <p>The <code>ID</code> of the state store</p>
id	<p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>HDFSBackedStateStore</code> state store is requested for the <code>textual representation</code></li> <li>• <code>StateStoreHandler</code> (of <code>SymmetricHashJoinStateManager</code>) is requested to <code>abortIfNeeded</code> and <code>getStateStore</code></li> </ul>
	<pre>iterator(): Iterator[UnsafeRowPair]</pre> <p>Returns an iterator with all the key-value pairs in the state store</p>
iterator	<p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>StateStoreRestoreExec</code> physical operator is requested to execute</li> <li>• <code>HDFSBackedStateStore</code> state store in particular and any <code>StateStore</code> in general are requested to <code>getRange</code></li> <li>• <code>StreamingAggregationStateManagerImplV1</code> state manager is requested for the <code>iterator</code> and <code>values</code></li> <li>• <code>StreamingAggregationStateManagerImplV2</code> state manager is requested to <code>iterator</code> and <code>values</code></li> </ul>
	<pre>metrics: StateStoreMetrics</pre> <p><code>StateStoreMetrics</code> of the state store</p>

metrics	<p>Used when:</p> <ul style="list-style-type: none"> <li>• StateStoreWriter stateful physical operator is requested to <a href="#">setStoreMetrics</a></li> <li>• StateStoreHandler (of <a href="#">SymmetricHashJoinStateManager</a>) is requested to <a href="#">commit</a> and for the metrics</li> </ul>
put	<pre>put(   key: UnsafeRow,   value: UnsafeRow): Unit</pre> <p>Stores (<i>puts</i>) the value for the (non-null) key</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <a href="#">StreamingDeduplicateExec</a> and <a href="#">StreamingGlobalLimitExec</a> physical operators are executed</li> <li>• <a href="#">StateManagerImplBase</a> is requested to <a href="#">putstate</a></li> <li>• <a href="#">StreamingAggregationStateManagerImplV1</a> and <a href="#">StreamingAggregationStateManagerImplV2</a> are requested to store a row in a state store</li> <li>• <a href="#">KeyToNumValuesStore</a> and <a href="#">KeyWithIndexToValueStore</a> are requested to store a new value for a given key</li> </ul>
remove	<pre>remove(key: UnsafeRow): Unit</pre> <p>Removes the (non-null) key from the state store</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• Physical operators with <a href="#">WatermarkSupport</a> are requested to <a href="#">removeKeysOlderThanWatermark</a></li> <li>• <a href="#">StateManagerImplBase</a> is requested to <a href="#">removeState</a></li> <li>• <a href="#">StreamingAggregationStateManagerBaseImpl</a> is requested to <a href="#">remove a key from a state store</a></li> <li>• <a href="#">KeyToNumValuesStore</a> is requested to <a href="#">remove a key</a></li> <li>• <a href="#">KeyWithIndexToValueStore</a> is requested to <a href="#">remove a key</a> and <a href="#">removeAllValues</a></li> </ul>
version	<pre>version: Long</pre> <p>Version of the state store</p>

Used exclusively when `HDFSBackedStateStore` state store is requested for a [new version](#) (that simply the current version incremented)

**Note**

`StateStore` was introduced in [\[SPARK-13809\]](#)[\[SQL\]](#) State store for streaming aggregations.

Read the motivation and design in [State Store for Streaming Aggregations](#).

**Tip**

Enable `ALL` logging level for `org.apache.spark.sql.execution.streaming.state.StateStore$` logger to see what happens inside.

Add the following line to `conf/log4j.properties`:

```
log4j.logger.org.apache.spark.sql.execution.streaming.state.StateStore$=ALL
```

Refer to [Logging](#).

## Creating (and Caching) RPC Endpoint Reference to StateStoreCoordinator for Executors — `coordinatorRef` Internal Object Method

```
coordinatorRef: Option[StateStoreCoordinatorRef]
```

`coordinatorRef` requests the `SparkEnv` helper object for the current `SparkEnv`.

If the `SparkEnv` is available and the `_coordRef` is not assigned yet, `coordinatorRef` prints out the following DEBUG message to the logs followed by requesting the `StateStoreCoordinatorRef` for the [StateStoreCoordinator endpoint](#).

```
Getting StateStoreCoordinatorRef
```

If the `SparkEnv` is available, `coordinatorRef` prints out the following INFO message to the logs:

```
Retrieved reference to StateStoreCoordinator: [_coordRef]
```

**Note**

`coordinatorRef` is used when `StateStore` helper object is requested to [reportActiveStoreInstance](#) (when `StateStore` object helper is requested to [find the StateStore by StateStoreProviderId](#)) and [verifyIfStoreInstanceActive](#) (when `StateStore` object helper is requested to [doMaintenance](#)).

## Unloading State Store Provider — `unload` Method

```
unload(storeProviderId: StateStoreProviderId): Unit
```

`unload` ...FIXME

Note

`unload` is used when `StateStore` helper object is requested to [stop](#) and [doMaintenance](#).

## stop Object Method

```
stop(): Unit
```

`stop` ...FIXME

Note

`stop` seems only be used in tests.

## Announcing New StateStoreProvider — `reportActiveStoreInstance` Internal Object Method

```
reportActiveStoreInstance(  
    storeProviderId: StateStoreProviderId): Unit
```

`reportActiveStoreInstance` takes the current host and `executorId` (from the `BlockManager` on the Spark executor) and requests the [StateStoreCoordinatorRef](#) to [reportActiveInstance](#).

Note

`reportActiveStoreInstance` uses `SparkEnv` to access the `BlockManager`.

In the end, `reportActiveStoreInstance` prints out the following INFO message to the logs:

```
Reported that the loaded instance [storeProviderId] is active
```

Note

`reportActiveStoreInstance` is used exclusively when `StateStore` utility is requested to [find the StateStore by StateStoreProviderId](#).

## MaintenanceTask Daemon Thread

`MaintenanceTask` is a daemon thread that [triggers maintenance work of registered StateStoreProviders](#).

When an error occurs, `MaintenanceTask` clears `loadedProviders` internal registry.

`MaintenanceTask` is scheduled on **state-store-maintenance-task** thread pool that runs periodically every `spark.sql.streaming.stateStore.maintenanceInterval` (default: `60s` ).

## Looking Up StateStore by Provider ID — `get` Object Method

```
get(
  storeProviderId: StateStoreProviderId,
  keySchema: StructType,
  valueSchema: StructType,
  indexOrdinal: Option[Int],
  version: Long,
  storeConf: StateStoreConf,
  hadoopConf: Configuration): StateStore
```

`get` finds `StateStore` for the specified `StateStoreProviderId` and `version`.

**Note**

The `version` is either the [current epoch](#) (in [Continuous Stream Processing](#)) or the [current batch ID](#) (in [Micro-Batch Stream Processing](#)).

Internally, `get` looks up the `StateStoreProvider` (by `storeProviderId`) in the `loadedProviders` internal cache. If unavailable, `get` uses the `StateStoreProvider` utility to create and initialize one.

`get` will also [start the periodic maintenance task](#) (unless already started) and announce the new `StateStoreProvider`.

In the end, `get` requests the `StateStoreProvider` to look up the `StateStore` by the specified `version`.

**Note**

- `get` is used when:
  - `StateStoreRDD` is requested to [compute a partition](#)
  - `StateStoreHandler` (of [SymmetricHashJoinStateManager](#)) is requested to [look up a StateStore](#) (by key and value schemas)

## Starting Periodic Maintenance Task (Unless Already Started) — `startMaintenanceIfNeeded` Internal Object Method

```
startMaintenanceIfNeeded(): Unit
```

`startMaintenanceIfNeeded` schedules `MaintenanceTask` to start after and every `spark.sql.streaming.stateStore.maintenanceInterval` (defaults to `60s` ).

**Note**

`startMaintenanceIfNeeded` does nothing when the maintenance task has already been started and is still running.

**Note**

`startMaintenanceIfNeeded` is used exclusively when `statestore` is requested to find the StateStore by `StateStoreProviderId`.

## Doing State Maintenance of Registered State Store Providers — `doMaintenance` Internal Object Method

```
doMaintenance(): Unit
```

Internally, `doMaintenance` prints the following DEBUG message to the logs:

```
Doing maintenance
```

`doMaintenance` then requests every `StateStoreProvider` (registered in `loadedProviders`) to do its own internal maintenance (only when a `StateStoreProvider` is still active).

When a `StateStoreProvider` is inactive, `doMaintenance` removes it from the provider registry and prints the following INFO message to the logs:

```
Unloaded [provider]
```

**Note**

`doMaintenance` is used exclusively in `MaintenanceTask` daemon thread.

## `verifyIfStoreInstanceActive` Internal Object Method

```
verifyIfStoreInstanceActive(storeProviderId: StateStoreProviderId): Boolean
```

```
verifyIfStoreInstanceActive ...FIXME
```

**Note**

`verifyIfStoreInstanceActive` is used exclusively when `StateStore` helper object is requested to `doMaintenance` (from a running `MaintenanceTask` daemon thread).

## Internal Properties

Name	Description
loadedProviders	<b>Loaded providers</b> internal cache, i.e. <a href="#">StateStoreProviders</a> per <a href="#">StateStoreProviderId</a> Used in...FIXME
_coordRef	<a href="#">StateStoreCoordinator RPC endpoint</a> (a <code>RpcEndpointRef</code> to <a href="#">StateStoreCoordinator</a> ) Used in...FIXME

# StateStoreId — Unique Identifier of State Store

`StateStoreId` is a unique identifier of a [state store](#) with the following attributes:

- **Checkpoint Root Location** - the root directory for state checkpointing
- **Operator ID** - a unique ID of the stateful operator
- **Partition ID** - the index of the partition
- **Store Name** - the name of the [state store](#) (default: `default`)

`StateStoreId` is [created](#) when:

- `StateStoreRDD` is requested for the [preferred locations of a partition](#) (executed on the driver) and to [compute it](#) (later on an executor)
- `StateStoreProviderId` helper object is requested to create a `StateStoreProviderId` (with a `StateStoreId` and the run ID of a streaming query) that is then used for the [preferred locations of a partition](#) of a `StateStoreAwareZipPartitionsRDD` (executed on the driver) and to...FIXME

The name of the **default state store** (for reading state store data that was generated before store names were used, i.e. in Spark 2.2 and earlier) is **default**.

## State Checkpoint Base Directory of Stateful Operator — `storeCheckpointLocation` Method

```
storeCheckpointLocation(): Path
```

`storeCheckpointLocation` is Hadoop DFS's [Path](#) of the checkpoint location (for the stateful operator by [operator ID](#), the partition by the [partition ID](#) in the [checkpoint root location](#)).

If the [default store name](#) is used (for Spark 2.2 and earlier), the [storeName](#) is not included in the path.

Note	<code>storeCheckpointLocation</code> is used exclusively when <code>HDFSBackedStateStoreProvider</code> is requested for the <a href="#">state checkpoint base directory</a> .
------	--

# HDFSBackedStateStore — State Store on HDFS-Compatible File System

`HDFSBackedStateStore` is a concrete [StateStore](#) that uses a Hadoop DFS-compatible file system for versioned state persistence.

`HDFSBackedStateStore` is [created](#) exclusively when `HDFSBackedStateStoreProvider` is requested for the [specified version of state \(store\)](#) for [update](#) (when `stateStore` utility is requested to [look up a StateStore by provider id](#)).

`HDFSBackedStateStore` uses the [StateStoreId](#) of the owning [HDFSBackedStateStoreProvider](#).

When requested for the textual representation, `HDFSBackedStateStore` gives `HDFSStateStore[id=(op=[operatorId],part=[partitionId]),dir=[baseDir]]`.

Tip

`HDFSBackedStateStore` is an internal class of `HDFSBackedStateStoreProvider` and uses its [logger](#).

## Creating HDFSBackedStateStore Instance

`HDFSBackedStateStore` takes the following to be created:

- Version
- State Map ( `ConcurrentHashMap[UnsafeRow, UnsafeRow]` )

`HDFSBackedStateStore` initializes the [internal properties](#).

## Internal State — `state` Internal Property

`state: STATE`

`state` is the current state of `HDFSBackedStateStore` and can be in one of the three possible states: [ABORTED](#), [COMMITTED](#), and [UPDATING](#).

State changes (to the internal `mapToUpdate` registry) are allowed as long as `HDFSBackedStateStore` is in the default [UPDATING](#) state. Right after a `HDFSBackedStateStore` transitions to either [COMMITTED](#) or [ABORTED](#) state, no further state changes are allowed.

Note	Don't get confused with the term "state" as there are two states: the internal state of <code>HDFSBackedStateStore</code> and the state of a streaming query (that <code>HDFSBackedStateStore</code> is responsible for).
------	---

Table 1. Internal States

Name	Description
ABORTED	After <code>abort</code>
COMMITTED	After <code>commit</code>  <code>hasCommitted</code> flag indicates whether <code>HDFSBackedStateStore</code> is in this state or not.
UPDATING	(default) Initial state after the <code>HDFSBackedStateStore</code> was created  Allows for state changes (e.g. <code>put</code> , <code>remove</code> , <code>getRange</code> ) and eventually committing or aborting them

## writeUpdateToDeltaFile Internal Method

```
writeUpdateToDeltaFile(
    output: DataOutputStream,
    key: UnsafeRow,
    value: UnsafeRow): Unit
```

Caution

FIXME

## put Method

```
put(
    key: UnsafeRow,
    value: UnsafeRow): Unit
```

Note

`put` is a part of [StateStore Contract](#) to...FIXME

`put` stores the copies of the key and value in `mapToUpdate` internal registry followed by writing them to a delta file (using `tempDeltaFileStream`).

`put` reports an `IllegalStateException` when `HDFSBackedStateStore` is not in [UPDATING](#) state:

Cannot put after already committed or aborted

## Committing State Changes — `commit` Method

`commit(): Long`

**Note** `commit` is part of the [StateStore Contract](#) to commit state changes.

`commit` requests the parent `HDFSBackedStateStoreProvider` to commit state changes (as a new version of state) (with the `newVersion`, the `mapToUpdate` and the compressed stream).

`commit` transitions `HDFSBackedStateStore` to [COMMITTED](#) state.

`commit` prints out the following INFO message to the logs:

Committed version [newVersion] for [this] to file [finalDeltaFile]

`commit` returns a `newVersion`.

`commit` throws an `IllegalStateException` when `HDFSBackedStateStore` is not in [UPDATING](#) state:

Cannot commit after already committed or aborted

`commit` throws an `IllegalStateException` for any `NonFatal` exception:

Error committing version [newVersion] into [this]

## Aborting State Changes — `abort` Method

`abort(): Unit`

**Note** `abort` is part of the [StateStore Contract](#) to abort the state changes.

`abort` ...FIXME

## Performance Metrics — `metrics` Method

`metrics: StateStoreMetrics`

Note	<code>metrics</code> is part of the <a href="#">StateStore Contract</a> to get the <a href="#">StateStoreMetrics</a> .
------	--

`metrics` requests the [performance metrics](#) of the parent [HDFSBackedStateStoreProvider](#).

The performance metrics of the provider used are only the ones listed in [supportedCustomMetrics](#).

In the end, `metrics` returns a new [StateStoreMetrics](#) with the following:

- [Total number of keys](#) as the size of `mapToUpdate`
- Memory used (in bytes) as the `memoryUsedBytes` metric (of the parent provider)
- [StateStoreCustomMetrics](#) as the `supportedCustomMetrics` and the `metricStateOnCurrentVersionSizeBytes` metric of the parent provider

## Are State Changes Committed? — `hasCommitted` Method

	<code>hasCommitted: Boolean</code>
--	------------------------------------

Note	<code>hasCommitted</code> is part of the <a href="#">StateStore Contract</a> to indicate whether state changes have been committed or not.
------	--

`hasCommitted` returns `true` when [HDFSBackedStateStore](#) is in [COMMITTED](#) state and `false` otherwise.

## Internal Properties

Name	Description
compressedStream	<pre>compressedStream: DataOutputStream</pre> <p>The compressed <a href="#">java.io.DataOutputStream</a> for the <a href="#">deltaFileStream</a></p>
deltaFileStream	<pre>deltaFileStream: CheckpointFileManager.CancellableFSDaOutputStream</pre>
finalDeltaFile	<pre>finalDeltaFile: Path</pre> <p>The Hadoop <a href="#">Path</a> of the <a href="#">deltaFile</a> for the <a href="#">version</a></p>
newVersion	<pre>newVersion: Long</pre> <p>Used exclusively when <a href="#">HDFSBackedStateStore</a> is requested for the <a href="#">finalDeltaFile</a>, to <a href="#">commit</a> and <a href="#">abort</a></p>

# StateStoreProvider Contract — State Store Providers

`StateStoreProvider` is the abstraction of state store providers that manage state stores in Stateful Stream Processing (e.g. for persisting running aggregates in Streaming Aggregation) in stateful streaming queries.

**Note** `StateStoreProvider` utility uses `spark.sql.streaming.stateStore.providerClass` internal configuration property for the name of the class of the default `StateStoreProvider` implementation.

**Note** `HDFSBackedStateStoreProvider` is the default and only known `StateStoreProvider` in Spark Structured Streaming.

Table 1. StateStoreProvider Contract

Method	Description
<code>close</code>	<pre>close(): Unit</pre> <p>Closes the state store provider Used exclusively when <code>statestore</code> helper object is requested to <a href="#">unload a state store provider</a></p>
<code>doMaintenance</code>	<pre>doMaintenance(): Unit = {}</pre> <p>Optional state maintenance Used exclusively when <code>statestore</code> utility is requested to <a href="#">perform maintenance of registered state store providers</a> (on a separate <code>MaintenanceTask</code> daemon thread)</p>
<code>getStore</code>	<pre>getStore(     version: Long): StateStore</pre> <p>Finds the <code>StateStore</code> for the specified version Used exclusively when <code>statestore</code> utility is requested to <a href="#">look up the StateStore by a given provider ID</a></p>

init	<pre>init(   statestoreId: StatestoreId,   keySchema: StructType,   valueSchema: StructType,   keyIndexOrdinal: Option[Int],   storeConfs: StatestoreConf,   hadoopConf: Configuration): Unit</pre> <p>Initializes the state store provider</p> <p>Used exclusively when <code>StateStoreProvider</code> helper object is requested to <a href="#">create and initialize the StateStoreProvider</a> for a given <code>StatestoreId</code> (when <code>StateStore</code> helper object is requested to <a href="#">retrieve a StateStore by ID and version</a>)</p>
statestoreId	<p><code>statestoreId: StatestoreId</code></p> <p><code>StatestoreId</code> associated with the provider (at <a href="#">initialization</a>)</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>HDFSBackedStateStore</code> is requested for the <a href="#">unique id</a></li> <li>• <code>HDFSBackedStateStoreProvider</code> is <a href="#">created</a> and requested for the <a href="#">textual representation</a></li> </ul>
supportedCustomMetrics	<p><code>supportedCustomMetrics: Seq[StatestoreCustomMetric]</code></p> <p><code>StatestoreCustomMetrics</code> of the state store provider</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>StateStoreWriter</code> stateful physical operators are requested for the <code>stateStoreCustomMetrics</code> (when requested for the <code>metrics</code> and <code>getProgress</code>)</li> <li>• <code>HDFSBackedStateStore</code> is requested for the <code>metrics</code></li> </ul>

## Creating and Initializing StateStoreProvider — `createAndInit` Object Method

```
createAndInit(
  statestoreId: StatestoreId,
  keySchema: StructType,
  valueSchema: StructType,
  indexOrdinal: Option[Int],
  storeConf: StatestoreConf,
  hadoopConf: Configuration): StateStoreProvider
```

`createAndInit` creates a new [StateStoreProvider](#) (per `spark.sql.streaming.stateStore.providerClass` internal configuration property).

`createAndInit` requests the `StateStoreProvider` to [initialize](#).

**Note**

`createAndInit` is used exclusively when `stateStore` utility is requested for the [StateStore](#) by given provider ID and version.

# StateStoreProviderId — Unique Identifier of State Store Provider

`StateStoreProviderId` is a unique identifier of a [state store provider](#) with the following properties:

- [StateStoreId](#)
- Run ID of a streaming query ([java.util.UUID](#))

In other words, `StateStoreProviderId` is a [StateStoreId](#) with the [run ID](#) that is different every restart.

`StateStoreProviderId` is used by the following execution components:

- `StateStoreCoordinator` to track the [executors of state store providers](#) (on the driver)
- `StateStore` object to manage [state store providers](#) (on executors)

`StateStoreProviderId` is [created](#) (directly or using [apply](#) factory method) when:

- `StateStoreRDD` is requested for the [placement preferences of a partition](#) and to [compute a partition](#)
- `StateStoreAwareZipPartitionsRDD` is requested for the [preferred locations of a partition](#)
- `StateStoreHandler` is requested to [look up a state store](#)

## Creating StateStoreProviderId — `apply` Factory Method

```
apply(
  stateInfo: StatefulOperatorStateInfo,
  partitionIndex: Int,
  storeName: String): StateStoreProviderId
```

`apply` simply creates a [new StateStoreProviderId](#) for the [StatefulOperatorStateInfo](#), the partition and the store name.

Internally, `apply` requests the `StatefulOperatorStateInfo` for the [checkpoint directory](#) (aka `checkpointLocation`) and the [stateful operator ID](#) and creates a new [StateStoreId](#) (with the `partitionIndex` and `storeName`).

In the end, `apply` requests the `StatefulOperatorStateInfo` for the [run ID of a streaming query](#) and creates a [new StateStoreProviderId](#) (together with the run ID).

	<p><code>apply</code> is used when:</p> <ul style="list-style-type: none"><li>• <code>StateStoreAwareZipPartitionsRDD</code> is requested for the preferred locations of a partition</li><li>• <code>StateStoreHandler</code> is requested to look up a state store</li></ul>
Note	

# HDFSBackedStateStoreProvider — Hadoop DFS-based StateStoreProvider

`HDFSBackedStateStoreProvider` is a [StateStoreProvider](#) that uses a Hadoop DFS-compatible file system for [versioned state checkpointing](#).

`HDFSBackedStateStoreProvider` is the default `stateStoreProvider` per the `spark.sql.streaming.stateStore.providerClass` internal configuration property.

`HDFSBackedStateStoreProvider` is [created](#) and immediately requested to [initialize](#) when `StateStoreProvider` utility is requested to [create and initialize a StateStoreProvider](#). That is when `HDFSBackedStateStoreProvider` is given the `StateStoreId` that uniquely identifies the [state store](#) to use for a stateful operator and a partition.

`HDFSStateStoreProvider` uses [HDFSBackedStateStores](#) to manage state ([one per version](#)).

`HDFSBackedStateStoreProvider` manages versioned state in delta and snapshot files (and uses a [cache](#) internally for faster access to state versions).

`HDFSBackedStateStoreProvider` takes no arguments to be created.

Tip	<p>Enable <code>ALL</code> logging level for <code>org.apache.spark.sql.execution.streaming.state.HDFSBackedStateStoreProvider</code> logger happens inside.</p> <p>Add the following line to <code>conf/log4j.properties</code> :</p> <pre>log4j.logger.org.apache.spark.sql.execution.streaming.state.HDFSBackedStateStoreProvider=ALL</pre> <p>Refer to <a href="#">Logging</a>.</p>
-----	---

## Performance Metrics

Name (in web UI)	Description
memoryUsedBytes	Estimated size of the <a href="#">loadedMaps</a> internal registry
count of cache hit on states cache in provider	The number of times <a href="#">loading the specified version of state</a> was successful and found ( <i>hit</i> ) the requested state version in the <a href="#">loadedMaps</a> internal cache
count of cache miss on states cache in provider	The number of times <a href="#">loading the specified version of state</a> could not find ( <i>missed</i> ) the requested state version in the <a href="#">loadedMaps</a> internal cache
estimated size of state only on current version	Estimated size of the <a href="#">current state</a> (of the <a href="#">HDFSBackedStateStore</a> )

## State Checkpoint Base Directory — `baseDir` Lazy Internal Property

`baseDir: Path`

`baseDir` is the base directory (as Hadoop DFS's [Path](#)) for [state checkpointing](#) (for [delta](#) and [snapshot](#) state files).

`baseDir` is initialized lazily since it is not yet known when [HDFSBackedStateStoreProvider](#) is [created](#).

`baseDir` is initialized and created based on the [state checkpoint base directory](#) of the [StateStoreId](#) when [HDFSBackedStateStoreProvider](#) is requested to [initialize](#).

## StateStoreId — Unique Identifier of State Store

As a [StateStoreProvider](#), [HDFSBackedStateStoreProvider](#) is associated with a [StateStoreId](#) (which is a unique identifier of the [state store](#) for a stateful operator and a partition).

[HDFSBackedStateStoreProvider](#) is given the [StateStoreId](#) at [initialization](#) (as requested by the [StateStoreProvider](#) contract).

The [StateStoreId](#) is then used for the following:

- [HDFSBackedStateStore](#) is requested for the [id](#)
- [HDFSBackedStateStoreProvider](#) is requested for the [textual representation](#) and the [state checkpoint base directory](#)

## Textual Representation — `toString` Method

```
toString: String
```

## Note

`toString` is part of the [java.lang.Object](#) contract for the string representation of the object.

`HDFSBackedStateStoreProvider` uses the [StateStoreId](#) and the state checkpoint base [directory](#) for the textual representation:

```
HDFSStateStoreProvider[id = (op=[operatorId],part=[partitionId]),dir = [baseDir]]
```

## Loading Specified Version of State (Store) For Update — `getStore` Method

```
getStore(  
    version: Long): StateStore
```

## Note

`getStore` is part of the [StateStoreProvider Contract](#) for the [StateStore](#) for a specified version.

`getStore` creates a new empty state (`ConcurrentHashMap[UnsafeRow, UnsafeRow]`) and loads the specified version of state (from internal cache or snapshot and delta files) for versions greater than `0`.

In the end, `getStore` creates a new [HDFSBackedStateStore](#) for the specified version with the new state and prints out the following INFO message to the logs:

```
Retrieved version [version] of [this] for update
```

`getStore` throws an `IllegalArgumentException` when the specified version is less than `0` (negative):

```
Version cannot be less than 0
```

## `deltaFile` Internal Method

```
deltaFile(version: Long): Path
```

`deltaFile` simply returns the Hadoop Path of the `[version].delta` file in the state checkpoint base directory.

**Note**

`deltaFile` is used when:

- `HDFSBackedStateStore` is created (and creates the final delta file)
- `HDFSBackedStateStoreProvider` is requested to `updateFromDeltaFile`

## snapshotFile Internal Method

```
snapshotFile(version: Long): Path
```

`snapshotFile` simply returns the Hadoop Path of the `[version].snapshot` file in the state checkpoint base directory.

**Note**

`snapshotFile` is used when `HDFSBackedStateStoreProvider` is requested to `writeSnapshotFile` or `readSnapshotFile`.

## Listing All Delta And Snapshot Files In State Checkpoint Directory — fetchFiles Internal Method

```
fetchFiles(): Seq[StoreFile]
```

`fetchFiles` requests the `CheckpointFileManager` for all the files in the state checkpoint directory.

For every file, `fetchFiles` splits the name into two parts with `.` (dot) as a separator (files with more or less than two parts are simply ignored) and registers a new `StoreFile` for `snapshot` and `delta` files:

- For `snapshot` files, `fetchFiles` creates a new `StoreFile` with `isSnapshot` flag on (`true`)
- For `delta` files, `fetchFiles` creates a new `StoreFile` with `isSnapshot` flag off (`false`)

**Note**

`delta` files are only registered if there was no `snapshot` file for the version.

`fetchFiles` prints out the following WARN message to the logs for any other files:

```
Could not identify file [path] for [this]
```

In the end, `fetchFiles` sorts the `StoreFiles` based on their version, prints out the following DEBUG message to the logs, and returns the files.

```
Current set of files for [this]: [storeFiles]
```

**Note**

`fetchFiles` is used when `HDFSBackedStateStoreProvider` is requested to `doSnapshot` and `cleanup`.

## Initializing StateStoreProvider — `init` Method

```
init(
  statestoreId: StatestoreId,
  keySchema: StructType,
  valueSchema: StructType,
  indexOrdinal: Option[Int],
  storeConf: StatestoreConf,
  hadoopConf: Configuration): Unit
```

**Note**

`init` is part of the [StateStoreProvider Contract](#) to initialize itself.

`init` records the values of the input arguments as the `statestoreId`, `keySchema`, `valueSchema`, `storeConf`, and `hadoopConf` internal properties.

`init` requests the given `StatestoreConf` for the `spark.sql.streaming.maxBatchesToRetainInMemory` configuration property (that is then recorded in the `numberOfVersionsToRetainInMemory` internal property).

In the end, `init` requests the [CheckpointFileManager](#) to [create](#) the `baseDir` directory (with parent directories).

## Finding Snapshot File and Delta Files For Version — `filesForVersion` Internal Method

```
filesForVersion(
  allFiles: Seq[StoreFile],
  version: Long): Seq[StoreFile]
```

`filesForVersion` finds the latest snapshot version among the given `allFiles` files up to and including the given version (it may or may not be available).

If a snapshot file was found (among the given file up to and including the given version), `filesForVersion` takes all delta files between the version of the snapshot file (exclusive) and the given version (inclusive) from the given `allFiles` files.

**Note**

The number of delta files should be the given version minus the snapshot version.

If a snapshot file was not found, `filesForVersion` takes all delta files up to the given version (inclusive) from the given `allFiles` files.

In the end, `filesForVersion` returns a snapshot version (if available) and all delta files up to the given version (inclusive).

**Note**

`filesForVersion` is used when `HDFSBackedStateStoreProvider` is requested to `doSnapshot` and `cleanup`.

## State Maintenance (Snapshotting and Cleaning Up)

### — `doMaintenance` Method

```
doMaintenance(): Unit
```

**Note**

`doMaintenance` is part of the [StateStoreProvider Contract](#) for optional state maintenance.

`doMaintenance` simply does [state snapshotting](#) followed by [cleaning up](#) (removing old state files).

In case of any non-fatal errors, `doMaintenance` simply prints out the following WARN message to the logs:

```
Error performing snapshot and cleaning up [this]
```

## State Snapshotting (Rolling Up Delta Files into Snapshot File) — `doSnapshot` Internal Method

```
doSnapshot(): Unit
```

`doSnapshot` lists all delta and snapshot files in the state checkpoint directory (`files`) and prints out the following DEBUG message to the logs:

```
fetchFiles() took [time] ms.
```

`doSnapshot` returns immediately (and does nothing) when there are no delta and snapshot files.

`doSnapshot` takes the version of the latest file (`lastversion`).

`doSnapshot` finds the snapshot file and delta files for the version (among the files and for the last version).

`doSnapshot` looks up the last version in the [internal state cache](#).

When the last version was found in the cache and the number of delta files is above `spark.sql.streaming.stateStore.minDeltasForSnapshot` internal threshold, `doSnapshot` writes a compressed snapshot file for the last version.

In the end, `doSnapshot` prints out the following DEBUG message to the logs:

```
writeSnapshotFile() took [time] ms.
```

In case of non-fatal errors, `doSnapshot` simply prints out the following WARN message to the logs:

```
Error doing snapshots for [this]
```

#### Note

`doSnapshot` is used exclusively when `HDFSBackedStateStoreProvider` is requested to [do state maintenance \(state snapshotting and cleaning up\)](#).

## Cleaning Up (Removing Old State Files) — `cleanup` Internal Method

```
cleanup(): Unit
```

`cleanup` lists all delta and snapshot files in the state checkpoint directory (`files`) and prints out the following DEBUG message to the logs:

```
fetchFiles() took [time] ms.
```

`cleanup` returns immediately (and does nothing) when there are no delta and snapshot files.

`cleanup` takes the version of the latest state file (`lastVersion`) and decrements it by `spark.sql.streaming.minBatchesToRetain` configuration property (default: `100`) that gives the earliest version to retain (and all older state files to be removed).

`cleanup` requests the [CheckpointFileManager](#) to delete the path of every old state file.

`cleanup` prints out the following DEBUG message to the logs:

```
deleting files took [time] ms.
```

In the end, `cleanup` prints out the following INFO message to the logs:

```
Deleted files older than [version] for [this]: [filesToDelete]
```

In case of a non-fatal exception, `cleanup` prints out the following WARN message to the logs:

```
Error cleaning up files for [this]
```

Note	<code>cleanup</code> is used exclusively when <code>HDFSBackedStateStoreProvider</code> is requested for <a href="#">state maintenance (state snapshotting and cleaning up)</a> .
------	---

## Closing State Store Provider — `close` Method

```
close(): Unit
```

Note	<code>close</code> is part of the <a href="#">StateStoreProvider Contract</a> to close the state store provider.
------	--

```
close ...FIXME
```

## `getMetricsForProvider` Method

```
getMetricsForProvider(): Map[String, Long]
```

`getMetricsForProvider` returns the following [performance metrics](#):

- [memoryUsedBytes](#)
- [metricLoadedMapCacheHit](#)
- [metricLoadedMapCacheMiss](#)

Note	<code>getMetricsForProvider</code> is used exclusively when <code>HDFSBackedStateStore</code> is requested for <a href="#">performance metrics</a> .
------	--

## Supported StateStoreCustomMetrics

### — supportedCustomMetrics Method

```
supportedCustomMetrics: Seq[StateStoreCustomMetric]
```

Note

`supportedCustomMetrics` is part of the [StateStoreProvider Contract](#) for the [StateStoreCustomMetrics](#) of a state store provider.

`supportedCustomMetrics` includes the following [StateStoreCustomMetrics](#):

- `metricStateOnCurrentVersionSizeBytes`
- `metricLoadedMapCacheHit`
- `metricLoadedMapCacheMiss`

## Committing State Changes (As New Version of State)

### — commitUpdates Internal Method

```
commitUpdates(  
    newVersion: Long,  
    map: ConcurrentHashMap[UnsafeRow, UnsafeRow],  
    output: DataOutputStream): Unit
```

`commitUpdates` [finalizeDeltaFile](#) (with the given `DataOutputStream`) followed by [caching the new version of state](#) (with the given `newVersion` and the `map` state).

Note

`commitUpdates` is used exclusively when `HDFSBackedStateStore` is requested to [commit state changes](#).

## Loading Specified Version of State (from Internal Cache or Snapshot and Delta Files) — loadMap Internal Method

```
loadMap(  
    version: Long): ConcurrentHashMap[UnsafeRow, UnsafeRow]
```

`loadMap` firstly tries to find the state version in the `loadedMaps` internal cache and, if found, returns it immediately and increments the `loadedMapCacheHitCount` metric.

If the requested state version could not be found in the `loadedMaps` internal cache, `loadMap` prints out the following WARN message to the logs:

The state for version [version] doesn't exist in loadedMaps. Reading snapshot file and delta files if needed...Note that this is normal for the first batch of starting query.

`loadMap` increments the [loadedMapCacheMissCount](#) metric.

`loadMap` tries to load the state snapshot file for the version and, if found, puts the version of state in the internal cache and returns it.

If not found, `loadMap` tries to find the most recent state version by decrementing the requested version until one is found in the [loadedMaps](#) internal cache or [loaded from a state snapshot \(file\)](#).

`loadMap updateFromDeltaFile` for all the remaining versions (from the snapshot version up to the requested one). `loadMap` puts the final version of state in the internal cache (the closest snapshot and the remaining delta versions) and returns it.

In the end, `loadMap` prints out the following DEBUG message to the logs:

```
Loading state for [version] takes [elapsedMs] ms.
```

Note	<code>loadMap</code> is used exclusively when <code>HDFSBackedStateStoreProvider</code> is requested for the <a href="#">specified version of a state store for update</a> .
------	--

## Loading State Snapshot File For Specified Version — `readSnapshotFile` Internal Method

```
readSnapshotFile(  
    version: Long): Option[ConcurrentHashMap[UnsafeRow, UnsafeRow]]
```

`readSnapshotFile` creates the path of the snapshot file for the given `version`.

`readSnapshotFile` requests the [CheckpointFileManager](#) to open the snapshot file for reading and creates a decompressed [DataInputStream](#) (`input`).

`readSnapshotFile` reads the decompressed input stream until an EOF (that is marked as the integer `-1` in the stream) and inserts key and value rows in a state map (`ConcurrentHashMap[UnsafeRow, UnsafeRow]`):

- First integer is the size of a key (buffer) followed by the key itself (of the size). `readSnapshotFile` creates an `UnsafeRow` for the key (with the number of fields as indicated by the number of fields of the [key schema](#)).

- Next integer is the size of a value (buffer) followed by the value itself (of the size). `readSnapshotFile` creates an `UnsafeRow` for the value (with the number of fields as indicated by the number of fields of the [value schema](#)).

In the end, `readSnapshotFile` prints out the following INFO message to the logs and returns the key-value map.

```
Read snapshot file for version [version] of [this] from [fileToRead]
```

In case of `FileNotFoundException` `readSnapshotFile` simply returns `None` (to indicate no snapshot state file was available and so no state for the version).

`readSnapshotFile` throws an `IOException` for the size of a key or a value below `0`:

```
Error reading snapshot file [fileToRead] of [this]: [key|value] size cannot be [keySize|valueSize]
```

Note	<code>readSnapshotFile</code> is used exclusively when <code>HDFSBackedStateStoreProvider</code> is requested to <a href="#">load the specified version of state (from the internal cache or snapshot and delta files)</a> .
------	--

## Updating State with State Changes For Specified Version (per Delta File) — `updateFromDeltaFile` Internal Method

```
updateFromDeltaFile(  
    version: Long,  
    map: ConcurrentHashMap[UnsafeRow, UnsafeRow]): Unit
```

Note	<code>updateFromDeltaFile</code> is very similar code-wise to <code>readSnapshotFile</code> with the two main differences:
------	--

- `updateFromDeltaFile` is given the state map to update (while `readSnapshotFile` loads the state from a snapshot file)
- `updateFromDeltaFile` removes a key from the state map when the value (size) is `-1` (while `readSnapshotFile` throws an `IOException`)

The following description is almost an exact copy of `readSnapshotFile` just for completeness.

`updateFromDeltaFile` creates the path of the delta file for the requested `version`.

`updateFromDeltaFile` requests the [CheckpointFileManager](#) to open the delta file for reading and creates a decompressed [DataInputStream](#) (`input`).

`updateFromDeltaFile` reads the decompressed input stream until an EOF (that is marked as the integer `-1` in the stream) and inserts key and value rows in the given state map:

- First integer is the size of a key (buffer) followed by the key itself (of the size). `updateFromDeltaFile` creates an [UnsafeRow](#) for the key (with the number of fields as indicated by the number of fields of the [key schema](#)).
- Next integer is the size of a value (buffer) followed by the value itself (of the size). `updateFromDeltaFile` creates an [UnsafeRow](#) for the value (with the number of fields as indicated by the number of fields of the [value schema](#)) or removes the corresponding key from the state map (if the value size is `-1` )

Note	<code>updateFromDeltaFile</code> removes the key-value entry from the state map if the value (size) is <code>-1</code> .
------	--

In the end, `updateFromDeltaFile` prints out the following INFO message to the logs and returns the key-value map.

```
Read delta file for version [version] of [this] from [fileToRead]
```

`updateFromDeltaFile` throws an [IllegalStateException](#) in case of [FileNotFoundException](#) while opening the delta file for the specified version:

```
Error reading delta file [fileToRead] of [this]: [fileToRead] does not exist
```

Note	<code>updateFromDeltaFile</code> is used exclusively when <a href="#">HDFSBackedStateStoreProvider</a> is requested to <a href="#">load the specified version of state (from the internal cache or snapshot and delta files)</a> .
------	--

## Caching New Version of State

### — `putStateIntoStateCacheMap` Internal Method

```
putStateIntoStateCacheMap(  
    newVersion: Long,  
    map: ConcurrentHashMap[UnsafeRow, UnsafeRow]): Unit
```

`putStateIntoStateCacheMap` registers state for a given version, i.e. adds the `map` state under the `newVersion` key in the [loadedMaps](#) internal registry.

With the `numberOfVersionsToRetainInMemory` threshold as `0` or below, `putStateIntoStateCacheMap` simply removes all entries from the `loadedMaps` internal registry and returns.

`putStateIntoStateCacheMap` removes the oldest state version(s) in the `loadedMaps` internal registry until its size is at the `numberOfVersionsToRetainInMemory` threshold.

With the size of the `loadedMaps` internal registry is at the `numberOfVersionsToRetainInMemory` threshold, `putStateIntoStateCacheMap` does two more optimizations per `newVersion`

- It does not add the given state when the version of the oldest state is earlier (larger) than the given `newVersion`
- It removes the oldest state when older (smaller) than the given `newVersion`

Note	<code>putStateIntoStateCacheMap</code> is used when <code>HDFSBackedStateStoreProvider</code> is requested to <b>commit state (as a new version)</b> and <b>load the specified version of state (from the internal cache or snapshot and delta files)</b> .
------	---

## Writing Compressed Snapshot File for Specified Version — `writeSnapshotFile` Internal Method

```
writeSnapshotFile(  
    version: Long,  
    map: ConcurrentHashMap[UnsafeRow, UnsafeRow]): Unit
```

`writeSnapshotFile` **snapshotFile** for the given version.

`writeSnapshotFile` requests the `CheckpointFileManager` to **create the snapshot file** (with overwriting enabled) and **compress the stream**.

For every key-value `UnsafeRow` pair in the given map, `writeSnapshotFile` writes the size of the key followed by the key itself (as bytes). `writeSnapshotFile` then writes the size of the value followed by the value itself (as bytes).

In the end, `writeSnapshotFile` prints out the following INFO message to the logs:

```
Written snapshot file for version [version] of [this] at [targetFile]
```

In case of any `Throwable` exception, `writeSnapshotFile` `cancelDeltaFile` and re-throws the exception.

## Note

`writeSnapshotFile` is used exclusively when `HDFSBackedStateStoreProvider` is requested to [doSnapshot](#).

**compressStream Internal Method**

```
compressStream(  
    outputStream: DataOutputStream): DataOutputStream
```

`compressStream` creates a new `LZ4CompressionCodec` (based on the [SparkConf](#)) and requests it to create a `LZ4BlockOutputStream` with the given `DataOutputStream`.

In the end, `compressStream` creates a new `DataOutputStream` with the `LZ4BlockOutputStream`.

## Note

`compressStream` is used when...FIXME

**cancelDeltaFile Internal Method**

```
cancelDeltaFile(  
    compressedStream: DataOutputStream,  
    rawStream: CancellableFSDataOutputStream): Unit
```

`cancelDeltaFile` ...FIXME

## Note

`cancelDeltaFile` is used when...FIXME

**finalizeDeltaFile Internal Method**

```
finalizeDeltaFile(  
    output: DataOutputStream): Unit
```

`finalizeDeltaFile` simply writes `-1` to the given `DataOutputStream` (to indicate end of file) and closes it.

## Note

`finalizeDeltaFile` is used exclusively when `HDFSBackedStateStoreProvider` is requested to [commit state changes \(a new version of state\)](#).

**Lookup Table (Cache) of States By Version  
— loadedMaps Internal Method**

```
loadedMaps: TreeMap[  
  Long, // version  
  ConcurrentHashMap[UnsafeRow, UnsafeRow]] // state (as keys and values)
```

`loadedMaps` is a [java.util.TreeMap](#) of state versions sorted according to the reversed ordering of the versions (i.e. long numbers).

A new entry (a version and the state updates) can only be added when `HDFSBackedStateStoreProvider` is requested to [putStateIntoStateCacheMap](#) (and only when the [spark.sql.streaming.maxBatchesToRetainInMemory](#) internal configuration is above `0`).

`loadedMaps` is mainly used when `HDFSBackedStateStoreProvider` is requested to [load the specified version of state \(from the internal cache or snapshot and delta files\)](#). Positive hits (when a version could be found in the cache) is available as the [count of cache hit on states cache in provider](#) performance metric while misses are counted in the [count of cache miss on states cache in provider](#) performance metric.

Note	With no or missing versions in cache <a href="#">count of cache miss on states cache in provider</a> metric should be above <code>0</code> while <a href="#">count of cache hit on states cache in provider</a> always <code>0</code> (or smaller than the other metric).
------	---

The estimated size of `loadedMaps` is available as the [memoryUsedBytes](#) performance metric.

The [spark.sql.streaming.maxBatchesToRetainInMemory](#) internal configuration is used as the threshold of the number of elements in `loadedMaps`. When `0` or negative, every [putStateIntoStateCacheMap](#) removes all elements in ([clears](#)) `loadedMaps`.

Note	It is possible to change the configuration at restart of a structured query.
------	--

The state deltas (the values) in `loadedMaps` are cleared (all entries removed) when `HDFSBackedStateStoreProvider` is requested to [close](#).

Used when `HDFSBackedStateStoreProvider` is requested for the following:

- [Cache a version of state](#)
- [Loading the specified version of state \(from the internal cache or snapshot and delta files\)](#)

## Internal Properties

Name	Description
fm	<p><code>CheckpointFileManager</code> for the state checkpoint base directory (and the Hadoop Configuration)</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• Creating a new <code>HDFSBackedStateStore</code> (to create the <code>CancellableFSDataOutputStream</code> for the <code>finalDeltaFile</code>)</li> <li>• <code>HDFSBackedStateStoreProvider</code> is requested to <code>initialize</code> (to create the state checkpoint base directory), <code>updateFromDeltaFile</code>, write the compressed snapshot file for a specified state version, <code>readSnapshotFile</code>, clean up, and list all delta and snapshot files in the state checkpoint directory</li> </ul>
hadoopConf	<p>Hadoop Configuration of the <code>CheckpointFileManager</code></p> <p>Given when <code>HDFSBackedStateStoreProvider</code> is requested to <code>initialize</code></p>
keySchema	<pre>keySchema: StructType</pre> <p>Schema of the state keys</p>
valueSchema	<pre>valueSchema: StructType</pre> <p>Schema of the state values</p>
numberOfVersionsToRetainInMemory	<pre>numberOfVersionsToRetainInMemory: Int</pre> <p><code>numberOfVersionsToRetainInMemory</code> is the maximum number of entries in the <code>loadedMaps</code> internal registry and is configured by the <code>spark.sql.streaming.maxBatchesToRetainInMemory</code> internal configuration.</p> <p><code>numberOfVersionsToRetainInMemory</code> is a threshold when <code>HDFSBackedStateStoreProvider</code> removes the last key from the <code>loadedMaps</code> internal registry (per reverse ordering of state versions) when requested to <code>putStateIntoStateCacheMap</code>.</p>
sparkConf	SparkConf



# StateStoreCoordinator RPC Endpoint — Tracking Locations of StateStores for StateStoreRDD

`StateStoreCoordinator` keeps track of [state stores](#) on Spark executors (per host and executor ID).

`StateStoreCoordinator` is used by `stateStoreRDD` when requested to [get the location preferences of partitions](#) (based on the location of the stores).

`StateStoreCoordinator` is a `ThreadSafeRpcEndpoint` RPC endpoint that manipulates [instances](#) registry through [RPC messages](#).

Table 1. StateStoreCoordinator RPC Endpoint's Messages and Message Handler

Message	Message Handler
<code>DeactivateInstances</code>	<p>Removes <a href="#">StateStoreProviderIds</a> of a streaming query (given <code>queryRunId</code> and <code>runId</code>) from the internal registry.</p> <p>Internally, <code>StateStoreCoordinator</code> finds the <code>StateStoreProvider</code> query per <code>queryRunId</code> and the given <code>runId</code> and removes them from the internal registry.</p> <p><code>StateStoreCoordinator</code> prints out the following DEBUG message:</p> <pre>Deactivating instances related to checkpoint location [runId]</pre>
<code>GetLocation</code>	<p>Gives the location of <a href="#">StateStoreProviderId</a> (from <a href="#">instances</a>) with the executor id on that host.</p> <p>You should see the following DEBUG message in the logs:</p> <pre>Got location of the state store [id]: [executorId]</pre>
<code>ReportActiveInstance</code>	<p>One-way asynchronous (fire-and-forget) message to register a <a href="#">StateStoreProviderId</a> on an executor (given <code>host</code> and <code>executorId</code>).</p> <p>Sent out exclusively when <code>StateStoreCoordinatorRef</code> RPC endpoint is requested to <a href="#">reportActiveInstance</a> (when <code>StateStore</code> utility is initialized by provider ID when the <code>StateStore</code> and a corresponding <code>StateStoreProvider</code> were just created and initialized).</p>

	<p>Internally, <code>StateStoreCoordinator</code> prints out the following DEBUG message:</p> <pre>Reported state store [id] is active at [executorId]</pre> <p>In the end, <code>StateStoreCoordinator</code> adds the <code>StateStoreProvider</code> to its internal registry.</p>
<code>stopCoordinator</code>	<p>Stops <code>StateStoreCoordinator</code> RPC Endpoint</p> <p>You should see the following DEBUG message in the logs:</p> <pre>StateStoreCoordinator stopped</pre>
<code>VerifyIfInstanceActive</code>	<p>Verifies if a given <code>StateStoreProviderId</code> is registered (in <code>instances</code>)</p> <p>You should see the following DEBUG message in the logs:</p> <pre>Verified that state store [id] is active: [response]</pre>
<b>Tip</b>	<p>Enable ALL logging level for <code>org.apache.spark.sql.execution.streaming.state.StateStoreCoordinator</code> logger to see what happens inside.</p> <p>Add the following line to <code>conf/log4j.properties</code>:</p> <pre>log4j.logger.org.apache.spark.sql.execution.streaming.state.StateStoreCoordinator=ALL</pre> <p>Refer to <a href="#">Logging</a>.</p>

## instances Internal Registry

```
instances: HashMap[StateStoreProviderId, ExecutorCacheTaskLocation]
```

`instances` is an internal registry of `StateStoreProviders` by their `StateStoreProviderIds` and `ExecutorCacheTaskLocations` (with a `host` and a `executorId`).

- A new `StateStoreProviderId` added when `StateStoreCoordinator` is requested to handle a [ReportActiveInstance](#) message
- All `StateStoreProviderIds` of a streaming query are removed when `StateStoreCoordinator` is requested to handle a [DeactivateInstances](#) message



# StateStoreCoordinatorRef — RPC Endpoint Reference to StateStoreCoordinator

`StateStoreCoordinatorRef` is used to (let the tasks on Spark executors to) send [messages](#) to the [StateStoreCoordinator](#) (that lives on the driver).

`StateStoreCoordinatorRef` is given the `RpcEndpointRef` to the [StateStoreCoordinator](#) RPC endpoint when created.

`StateStoreCoordinatorRef` is created through `StateStoreCoordinatorRef` helper object when requested to create one for the [driver](#) (when `StreamingQueryManager` is created) or an [executor](#) (when `Statestore` helper object is requested for the [RPC endpoint reference to StateStoreCoordinator for Executors](#)).

Table 1. StateStoreCoordinatorRef's Methods and Underlying RPC Messages

Method	Description
<code>deactivateInstances</code>	<pre>deactivateInstances(runId: UUID): Unit</pre> <p>Requests the <a href="#">RpcEndpointRef</a> to send a <a href="#">DeactivateInstances</a> synchronous message with the given <code>runId</code> and waits for a <code>true</code> / <code>false</code> response</p> <p>Used exclusively when <code>StreamingQueryManager</code> is requested to <a href="#">handle termination of a streaming query</a> (when <code>StreamExecution</code> is requested to <a href="#">run a streaming query</a> and the query <a href="#">has finished (running streaming batches)</a>).</p>
<code>getLocation</code>	<pre>getLocation(     stateStoreProviderId: StateStoreProviderId): Option[String]</pre> <p>Requests the <a href="#">RpcEndpointRef</a> to send a <a href="#">GetLocation</a> synchronous message with the given <a href="#">StateStoreProviderId</a> and waits for the location</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>StateStoreAwareZipPartitionsRDD</code> is requested for the <a href="#">preferred locations of a partition</a> (when <code>StreamingSymmetricHashJoinExec</code> physical operator is requested to <a href="#">execute and generate a recipe for a distributed computation (as an RDD[InternalRow])</a>)</li> <li>• <code>StateStoreRDD</code> is requested for <a href="#">preferred locations for a task for a partition</a></li> </ul>

	<pre>reportActiveInstance(     stateStoreProviderId: StateStoreProviderId,     host: String,     executorId: String): Unit</pre>
reportActiveInstance	<p>Requests the <a href="#">RpcEndpointRef</a> to send a <a href="#">ReportActiveInstance</a> one-way asynchronous (fire-and-forget) message with the given <a href="#">StateStoreProviderId</a>, <a href="#">host</a> and <a href="#">executorId</a></p> <p>Used exclusively when <a href="#">statestore</a> utility is requested for <a href="#">reportActiveStoreInstance</a> (when <a href="#">statestore</a> utility is requested to look up the <a href="#">StateStore</a> by <a href="#">StateStoreProviderId</a>)</p>
stop	<pre>stop(): Unit</pre> <p>Requests the <a href="#">RpcEndpointRef</a> to send a <a href="#">StopCoordinator</a> synchronous message</p> <p>Used exclusively for unit testing</p>
verifyIfInstanceActive	<pre>verifyIfInstanceActive(     stateStoreProviderId: StateStoreProviderId,     executorId: String): Boolean</pre> <p>Requests the <a href="#">RpcEndpointRef</a> to send a <a href="#">VerifyIfInstanceActive</a> synchronous message with the given <a href="#">StateStoreProviderId</a> and <a href="#">executorId</a>, and waits for a <a href="#">true</a> / <a href="#">false</a> response</p> <p>Used exclusively when <a href="#">statestore</a> helper object is requested for <a href="#">verifyIfStoreInstanceActive</a> (when requested to <a href="#">doMaintenance</a> from a running <a href="#">MaintenanceTask</a> daemon thread)</p>

## Creating StateStoreCoordinatorRef to StateStoreCoordinator RPC Endpoint for Driver — `forDriver` Factory Method

```
forDriver(env: SparkEnv): StateStoreCoordinatorRef
```

`forDriver` ...FIXME

Note	<code>forDriver</code> is used exclusively when <a href="#">StreamingQueryManager</a> is created.
------	---

## Creating StateStoreCoordinatorRef to StateStoreCoordinator RPC Endpoint for Executor — forExecutor Factory Method

```
forExecutor(env: SparkEnv): StateStoreCoordinatorRef
```

```
forExecutor ...FIXME
```

Note

forExecutor is used exclusively when StateStore helper object is requested for the [RPC endpoint reference to StateStoreCoordinator for Executors](#).

# WatermarkSupport Contract — Unary Physical Operators with Streaming Watermark Support

`WatermarkSupport` is the [abstraction](#) of unary physical operators (`UnaryExecNode`) with support for streaming event-time watermark.

Note

**Watermark** (aka "allowed lateness") is a moving threshold of **event time** and specifies what data to consider for aggregations, i.e. the threshold of late data so the engine can automatically drop incoming late data given event time and clean up old state accordingly.

Read the official documentation of Spark in [Handling Late Data and Watermarking](#).

Table 1. WatermarkSupport's (Lazily-Initialized) Properties

Property	Description					
	<p>Optional Catalyst expression that matches rows older than the event time watermark.</p> <table border="1"> <tr> <td>Note</td><td>Use <a href="#">withWatermark</a> operator to specify streaming watermark.</td></tr> </table>		Note	Use <a href="#">withWatermark</a> operator to specify streaming watermark.		
Note	Use <a href="#">withWatermark</a> operator to specify streaming watermark.					
<code>watermarkExpression</code>	<p>When initialized, <code>watermarkExpression</code> finds <code>spark.watermark</code> watermark attribute in the child output's metadata.</p> <p>If found, <code>watermarkExpression</code> creates <code>evictionExpression</code> attribute that is less than or equal <code>eventTimeWatermark</code>.</p> <p>The <code>watermark</code> attribute may be of type <code>StructType</code>. If it is, <code>watermarkExpression</code> uses the first field as the watermark.</p> <p><code>watermarkExpression</code> prints out the following INFO message if <code>spark.watermarkDelayMs</code> watermark attribute is found.</p> <pre>INFO [physicaloperator]Exec: Filtering state store on: [</pre> <table border="1"> <tr> <td>Note</td><td><code>physicaloperator</code> can be <code>FlatMapGroupsWithState</code>, <code>StateStoreSaveExec</code> or <code>StreamingDeduplicateExec</code>.</td></tr> </table> <table border="1"> <tr> <td>Tip</td><td>Enable INFO logging level for one of the stateful physical operators to see the INFO message in the logs.</td></tr> </table>		Note	<code>physicaloperator</code> can be <code>FlatMapGroupsWithState</code> , <code>StateStoreSaveExec</code> or <code>StreamingDeduplicateExec</code> .	Tip	Enable INFO logging level for one of the stateful physical operators to see the INFO message in the logs.
Note	<code>physicaloperator</code> can be <code>FlatMapGroupsWithState</code> , <code>StateStoreSaveExec</code> or <code>StreamingDeduplicateExec</code> .					
Tip	Enable INFO logging level for one of the stateful physical operators to see the INFO message in the logs.					
<code>watermarkPredicateForData</code>	Optional <code>Predicate</code> that uses <code>watermarkExpression</code> and the <code>keyExpressions</code> to match rows older than the event-time watermark.					
<code>watermarkPredicateForKeys</code>	Optional <code>Predicate</code> that uses <code>keyExpressions</code> to match rows older than the event time watermark.					

## WatermarkSupport Contract

```
package org.apache.spark.sql.execution.streaming

trait WatermarkSupport extends UnaryExecNode {
    // only required methods that have no implementation
    def eventTimeWatermark: Option[Long]
    def keyExpressions: Seq[Attribute]
}
```

Table 2. WatermarkSupport Contract

Method	Description
eventTimeWatermark	Used mainly in <a href="#">watermarkExpression</a> to create a <code>LessThanOrEqual</code> Catalyst binary expression that matches rows older than the watermark.
keyExpressions	<p>Grouping keys (in <a href="#">FlatMapGroupsWithStateExec</a>), duplicate keys (in <a href="#">StreamingDeduplicateExec</a>) or key attributes (in <a href="#">StateStoreSaveExec</a>) with at most one that may have <code>spark.watermarkDelayMs</code> watermark attribute in metadata</p> <p>Used in <a href="#">watermarkPredicateForKeys</a> to create a <code>Predicate</code> to match rows older than the event time watermark.</p> <p>Used also when <a href="#">StateStoreSaveExec</a> and <a href="#">StreamingDeduplicateExec</a> physical operators are executed.</p>

## Removing Keys From StateStore Older Than Watermark

### — `removeKeysOlderThanWatermark` Method

```
removeKeysOlderThanWatermark(store: StateStore): Unit
```

`removeKeysOlderThanWatermark` requests the input `store` for all rows.

`removeKeysOlderThanWatermark` then uses [watermarkPredicateForKeys](#) to remove matching rows from the store.

#### Note

`removeKeysOlderThanWatermark` is used exclusively when `StreamingDeduplicateExec` physical operator is requested to execute and generate a recipe for a distributed computation (as an `RDD[InternalRow]`).

## removeKeysOlderThanWatermark Method

```
removeKeysOlderThanWatermark(
  storeManager: StreamingAggregationStateManager,
  store: StateStore): Unit
```

`removeKeysOlderThanWatermark` ...FIXME

Note	<p><code>removeKeysOlderThanWatermark</code> is used exclusively when <code>StateStoreSaveExec</code> physical operator is requested to execute and generate a recipe for a distributed computation (as an <code>RDD[InternalRow]</code>).</p>
------	--

# StatefulOperator Contract — Physical Operators That Read or Write to StateStore

`StatefulOperator` is the base of physical operators that read or write state (described by `stateInfo`).

Table 1. StatefulOperator Contract

Method	Description
<code>stateInfo</code>	<code>stateInfo: Option[StatefulOperatorStateInfo]</code>  The <code>StatefulOperatorStateInfo</code> of the physical operator

Table 2. StatefulOperators (Direct Implementations)

StatefulOperator	Description
<code>StateStoreReader</code>	
<code>StateStoreWriter</code>	Physical operator that writes to a state store and collects the write metrics for execution progress reporting

# StateStoreReader

StateStoreReader is...FIXME

# StateStoreWriter Contract — Stateful Physical Operators That Write to State Store

`StateStoreWriter` is the extension of the [StatefulOperator Contract](#) for physical operators that write to a state store and collect the [write metrics](#) for [execution progress reporting](#).

Table 1. StateStoreWriters

StateStoreWriter	Description
<a href="#">FlatMapGroupsWithStateExec</a>	
<a href="#">StateStoreSaveExec</a>	
<a href="#">StreamingDeduplicateExec</a>	
<a href="#">StreamingGlobalLimitExec</a>	
<a href="#">StreamingSymmetricHashJoinExec</a>	

## Performance Metrics (SQLMetrics)

Name (in web UI)	Description
number of output rows	
number of total state rows	
number of updated state rows	
total time to update rows	
total time to remove rows	
time to commit changes	
memory used by state	

## Setting StateStore-Specific Metrics for Stateful Physical Operator — `setStoreMetrics` Method

```
setStoreMetrics(store: StateStore): Unit
```

`setStoreMetrics` requests the specified `StateStore` for the `metrics` and records the following metrics of a physical operator:

- `numTotalStateRows` as the `number of keys`
- `stateMemory` as the `memory used (in bytes)`

`setStoreMetrics` records the `custom metrics`.

Note

`setStoreMetrics` is used when the following physical operators are executed:

- `FlatMapGroupsWithStateExec`
- `StateStoreSaveExec`
- `StreamingDeduplicateExec`
- `StreamingGlobalLimitExec`

## getProgress Method

```
getProgress(): StateOperatorProgress
```

`getProgress` ...FIXME

Note

`getProgress` is used exclusively when `ProgressReporter` is requested to `extractStateOperatorMetrics` (when `MicroBatchExecution` is requested to `run` the activated streaming query).

## Checking Out Whether Last Batch Execution Requires Another Non-Data Batch or Not — `shouldRunAnotherBatch` Method

```
shouldRunAnotherBatch(newMetadata: OffsetSeqMetadata): Boolean
```

`shouldRunAnotherBatch` is negative (`false`) by default (to indicate that another non-data batch is not required given the `OffsetSeqMetadata` with the event-time watermark and the batch timestamp).

**Note**

`shouldRunAnotherBatch` is used exclusively when `IncrementalExecution` is requested to check out whether the last batch execution requires another batch (when `MicroBatchExecution` is requested to run the activated streaming query).

## stateStoreCustomMetrics Internal Method

```
stateStoreCustomMetrics: Map[String, SQLMetric]
```

```
stateStoreCustomMetrics ...FIXME
```

**Note**

`stateStoreCustomMetrics` is used when `StateStoreWriter` is requested for the metrics and `getProgress`.

## timeTakenMs Method

```
timeTakenMs(body: => Unit): Long
```

```
timeTakenMs ...FIXME
```

**Note**

`timeTakenMs` is used when...FIXME

# StatefulOperatorStateInfo

`StatefulOperatorStateInfo` identifies the state store for a given stateful physical operator:

- Checkpoint directory (`checkpointLocation`)
- Run ID of a streaming query (`queryRunId`)
- Stateful operator ID (`operatorId`)
- State version (`storeVersion`)
- Number of partitions

`StatefulOperatorStateInfo` is created exclusively when `IncrementalExecution` is requested for `nextStatefulOperationStateInfo`.

When requested for a textual representation (`toString`), `StatefulOperatorStateInfo` returns the following:

```
state info [ checkpoint = [checkpointLocation], runId = [queryRunId], opId = [operatorId], ver = [storeVersion], numPartitions = [numPartitions] ]
```

## State Version and Batch ID

When created (when `IncrementalExecution` is requested for the next `StatefulOperatorStateInfo`), a `StatefulOperatorStateInfo` is given a state version.

The state version is exactly the batch ID of the `IncrementalExecution`.

# StateStoreMetrics

`StateStoreMetrics` holds the performance metrics of a [state store](#):

- Number of keys
- Memory used (in bytes)
- [StateStoreCustomMetrics](#) with their current values (`Map[StateStoreCustomMetric, Long]`)

`StateStoreMetrics` is used (and [created](#)) when the following are requested for the performance metrics:

- [StateStore](#)
- [StateStoreHandler](#)
- [SymmetricHashJoinStateManager](#)

# StateStoreCustomMetric Contract

`StateStoreCustomMetric` is the [abstraction of metrics](#) that a state store may wish to expose (as [StateStoreMetrics](#) or [supportedCustomMetrics](#)).

`StateStoreCustomMetric` is used when:

- `StateStoreProvider` is requested for the [custom metrics](#)
- `StateStoreMetrics` is [created](#)

Table 1. StateStoreCustomMetric Contract

Method	Description
<code>desc</code>	<p><code>desc: String</code></p> <p>Description of the custom metrics</p>
<code>name</code>	<p><code>name: String</code></p> <p>Name of the custom metrics</p>

Table 2. StateStoreCustomMetrics

StateStoreCustomMetric	Description
<code>StateStoreCustomSizeMetric</code>	
<code>StateStoreCustomSumMetric</code>	
<code>StateStoreCustomTimingMetric</code>	

# StateStoreUpdater

StateStoreUpdater is...FIXME

## updateStateForKeysWithData Method

Caution	FIXME
---------	-------

## updateStateForTimedOutKeys Method

Caution	FIXME
---------	-------

# EventTimeStatsAccum Accumulator — Event-Time Column Statistics for EventTimeWatermarkExec Physical Operator

`EventTimeStatsAccum` is a Spark accumulator that is used for the [statistics of the event-time column](#) (that `EventTimeWatermarkExec` physical operator uses for event-time watermark):

- Maximum value
- Minimum value
- Average value
- Number of updates (count)

`EventTimeStatsAccum` is [created](#) and registered exclusively for `EventTimeWatermarkExec` physical operator.

## Note

When `EventTimeWatermarkExec` physical operator is requested to [execute and generate a recipe for a distributed computation \(as a `RDD\[InternalRow\]`\)](#), every task simply [adds](#) the values of the event-time watermark column to the `EventTimeStatsAccum` accumulator.

As per design of Spark accumulators in Apache Spark, accumulator updates are automatically sent out (*propagated*) from tasks to the driver every heartbeat and then they are accumulated together.

## Tip

Read up on [Accumulators](#) in [The Internals of Apache Spark](#) book.

`EventTimeStatsAccum` takes a single `EventTimeStats` to be created (default: `zero`).

## Accumulating Value — `add` Method

```
add(v: Long): Unit
```

## Note

`add` is part of the `AccumulatorV2` Contract to add (*accumulate*) a given value.

`add` simply requests the `EventTimeStats` to [add](#) the given `v` value.

## Note

`add` is used exclusively when `EventTimeWatermarkExec` physical operator is requested to [execute and generate a recipe for a distributed computation \(as a `RDD\[InternalRow\]`\)](#).

## EventTimeStats

`EventTimeStats` is a Scala case class for the event-time column statistics.

`EventTimeStats` defines a special value `zero` with the following values:

- `Long.MinValue` for the `max`
- `Long.MaxValue` for the `min`
- `0.0` for the `avg`
- `0L` for the `count`

### EventTimeStats.add Method

```
add(eventTime: Long): Unit
```

`add` simply updates the event-time column statistics per given `eventTime`.

Note	<code>add</code> is used exclusively when <code>EventTimeStatsAccum</code> is requested to accumulate the value of an event-time column.
------	--

### EventTimeStats.merge Method

```
merge(that: EventTimeStats): Unit
```

`merge` ...FIXME

Note	<code>merge</code> is used when...FIXME
------	---

# StateStoreConf

StateStoreConf is...FIXME

Table 1. StateStoreConf's Properties

Name	Configuration Property
minDeltasForSnapshot	<code>spark.sql.streaming.stateStore.minDeltasForSnapshot</code>
maxVersionsToRetainInMemory	<code>spark.sql.streaming.maxBatchesToRetainInMemory</code>
minVersionsToRetain	<code>spark.sql.streaming.minBatchesToRetain</code> Used exclusively when <code>HDFSBackedStateStoreProvider</code> is requested for <code>cleanup</code> .
providerClass	<code>spark.sql.streaming.stateStore.providerClass</code> Used exclusively when <code>stateStoreProvider</code> helper object is requested to <code>create and initialize the StateStoreProvider</code> .

# Arbitrary Stateful Streaming Aggregation

**Arbitrary Stateful Streaming Aggregation** is a [streaming aggregation query](#) that uses the following [KeyValueGroupedDataset](#) operators:

- [mapGroupsWithState](#) for implicit state logic
- [flatMapGroupsWithState](#) for explicit state logic

`KeyValueGroupedDataset` represents a grouped dataset as a result of [Dataset.groupByKey](#) operator.

`mapGroupsWithState` and `flatMapGroupsWithState` operators use [GroupState](#) as **group streaming aggregation state** that is created separately for every **aggregation key** with an **aggregation state value** (of a user-defined type).

`mapGroupsWithState` and `flatMapGroupsWithState` operators use [GroupStateTimeout](#) as an **aggregation state timeout** that defines when a [GroupState](#) can be considered **timed-out (expired)**.

## Demos

Use the following demos and complete applications to learn more:

- [Demo: Internals of FlatMapGroupsWithStateExec Physical Operator](#)
- [Demo: Arbitrary Stateful Streaming Aggregation with KeyValueGroupedDataset.flatMapGroupsWithState Operator](#)
- [groupByKey Streaming Aggregation in Update Mode](#)
- [FlatMapGroupsWithStateApp](#)

## Performance Metrics

Arbitrary Stateful Streaming Aggregation uses **performance metrics** (of the [StateStoreWriter](#) through [FlatMapGroupsWithStateExec](#) physical operator).

## Internals

One of the most important internal execution components of Arbitrary Stateful Streaming Aggregation is [FlatMapGroupsWithStateExec](#) physical operator.

When requested to [execute and generate a recipe for a distributed computation](#) (as an [RDD\[InternalRow\]](#)), [FlatMapGroupsWithStateExec](#) first validates a selected [GroupStateTimeout](#):

- For [ProcessingTimeTimeout](#), batch timeout [threshold](#) has to be defined
- For [EventTimeTimeout](#), event-time watermark has to be defined and the [input schema has the watermark attribute](#)

Note	FIXME When are the above requirements met?
------	--

[FlatMapGroupsWithStateExec](#) physical operator then [mapPartitionsWithStateStore](#) with a custom [storeUpdateFunction](#) of the following signature:

```
(StateStore, Iterator[T]) => Iterator[U]
```

While generating the recipe, [FlatMapGroupsWithStateExec](#) uses [StateStoreOps](#) extension method object to register a listener that is executed on a task completion. The listener makes sure that a given [StateStore](#) has all state changes either [committed](#) or [aborted](#).

In the end, [FlatMapGroupsWithStateExec](#) creates a new [StateStoreRDD](#) and adds it to the RDD lineage.

[StateStoreRDD](#) is used to properly distribute tasks across executors (per [preferred locations](#)) with help of [StateStoreCoordinator](#) (that runs on the driver).

[StateStoreRDD](#) uses [StateStore](#) helper to look up a [StateStore](#) by [StateStoreProviderId](#) and store version.

[FlatMapGroupsWithStateExec](#) physical operator uses [state managers](#) that are different than [state managers](#) for [Streaming Aggregation](#). [StateStore](#) abstraction is the same as in [Streaming Aggregation](#).

One of the important execution steps is when [InputProcessor](#) (of [FlatMapGroupsWithStateExec](#) physical operator) is requested to [callFunctionAndUpdateState](#). That executes the **user-defined state function** on a per-group state key object, value objects, and a [GroupStateImpl](#).

# GroupState — Group State in Arbitrary Stateful Streaming Aggregation

`GroupState` is an abstraction of `group state` (of type `s`) in `Arbitrary Stateful Streaming Aggregation`.

`GroupState` is used with the following `KeyValueGroupedDataset` operations:

- `mapGroupsWithState`
- `flatMapGroupsWithState`

`GroupState` is created separately for every **aggregation key** to hold a state as an **aggregation state value**.

Table 1. GroupState Contract

Method	Description
<code>exists</code>	<pre>exists: Boolean</pre> <p>Checks whether the state value exists or not If not exists, <code>get</code> throws a <code>NoSuchElementException</code>. Use <code>getOption</code> instead.</p>
<code>get</code>	<pre>get: S</pre> <p>Gets the state value if it <code>exists</code> or throws a <code>NoSuchElementException</code></p>
<code>getCurrentProcessingTimeMs</code>	<pre>getCurrentProcessingTimeMs(): Long</pre> <p>Gets the current processing time (as milliseconds in epoch time)</p>
<code>getCurrentWatermarkMs</code>	<pre>getCurrentWatermarkMs(): Long</pre> <p>Gets the current event time watermark (as milliseconds in epoch time)</p>
	<pre>getOption: Option[S]</pre>

	<p>Gets the state value as a Scala <code>option</code> (regardless whether it <a href="#">exists</a> or not)</p> <p><code>getOption</code></p> <ul style="list-style-type: none"> <li>• <code>InputProcessor</code> is requested to <code>callFunctionAndUpdateState</code> (when the row iterator is consumed and a state value has been updated, removed or timeout changed)</li> <li>• <code>GroupStateImpl</code> is requested for the <a href="#">textual representation</a></li> </ul>
<code>hasTimedOut</code>	<pre>hasTimedOut: Boolean</pre> <p>Whether the state (for a given key) has timed out or not.</p> <p>Can only be <code>true</code> when timeouts are enabled using <a href="#">setTimeoutDuration</a></p>
<code>remove</code>	<pre>remove(): Unit</pre> <p>Removes the state</p>
<code>setTimeoutDuration</code>	<pre>setTimeoutDuration(durationMs: Long): Unit setTimeoutDuration(duration: String): Unit</pre> <p>Specifies the <b>timeout duration</b> for the state key (in millis or as a string, e.g. "10 seconds", "1 hour") for <a href="#">GroupStateTimeout.ProcessingTimeTimeout</a></p>
<code>setTimeoutTimestamp</code>	<pre>setTimeoutTimestamp(timestamp: java.sql.Date): Unit setTimeoutTimestamp(     timestamp: java.sql.Date,     additionalDuration: String): Unit setTimeoutTimestamp(timestampMs: Long): Unit setTimeoutTimestamp(     timestampMs: Long,     additionalDuration: String): Unit</pre> <p>Specifies the <b>timeout timestamp</b> for the state key for <a href="#">GroupStateTimeout.EventTimeTimeout</a></p>
<code>update</code>	<pre>update(newState: S): Unit</pre> <p>Updates the state (sets the state to a new value)</p>

Note	<p><a href="#">GroupStateImpl</a> is the default and only known implementation of the <a href="#">GroupState Contract</a> in Spark Structured Streaming.</p>
------	--

# GroupStateImpl

`GroupStateImpl` is the default and only known `GroupState` in Spark Structured Streaming.

`GroupStateImpl` holds per-group `state value` of type `s` per group key.

`GroupStateImpl` is `created` when `GroupStateImpl` helper object is requested for the following:

- `createForStreaming`
- `createForBatch`

## Creating GroupStateImpl Instance

`GroupStateImpl` takes the following to be created:

- State value (of type `s`)
- Batch processing time
- `eventTimeWatermarkMs`
- `GroupStateTimeout`
- `hasTimedOut` flag
- `watermarkPresent` flag

`GroupStateImpl` initializes the [internal properties](#).

## Creating GroupStateImpl for Streaming Query — `createForStreaming` Factory Method

```
createForStreaming[S](  
    optionalValue: Option[S],  
    batchProcessingTimeMs: Long,  
    eventTimeWatermarkMs: Long,  
    timeoutConf: GroupStateTimeout,  
    hasTimedOut: Boolean,  
    watermarkPresent: Boolean): GroupStateImpl[S]
```

`createForStreaming` simply creates a new `GroupStateImpl` with the given input arguments.

**Note**

`createForStreaming` is used exclusively when `InputProcessor` is requested to `callFunctionAndUpdateState` (when `InputProcessor` is requested to `processNewData` and `processTimedOutState`).

**Creating GroupStateImpl for Batch Query****— `createForBatch` Factory Method**

```
createForBatch(  
    timeoutConf: GroupStateTimeout,  
    watermarkPresent: Boolean): GroupStateImpl[Any]
```

`createForBatch` ...FIXME

**Note**

`createForBatch` is used when...FIXME

**Textual Representation — `toString` Method**

```
toString: String
```

**Note**

`toString` is part of the `java.lang.Object` contract for the string representation of the object.

`toString` ...FIXME

**Specifying Timeout Duration for ProcessingTimeTimeout****— `setTimeoutDuration` Method**

```
setTimeoutDuration(durationMs: Long): Unit
```

**Note**

`setTimeoutDuration` is part of the `GroupState Contract` to specify timeout duration for the state key (in millis or as a string).

`setTimeoutDuration` ...FIXME

**Specifying Timeout Timestamp for EventTimeTimeout****— `setTimeoutTimestamp` Method**

```
setTimeoutTimestamp(durationMs: Long): Unit
```

**Note**

`setTimeoutTimestamp` is part of the [GroupState Contract](#) to specify timeout timestamp for the state key.

`setTimeoutTimestamp ...FIXME`

## Getting Processing Time — `getCurrentProcessingTimeMs` Method

`getCurrentProcessingTimeMs(): Long`

**Note**

`getCurrentProcessingTimeMs` is part of the [GroupState Contract](#) to get the current processing time (as milliseconds in epoch time).

`getCurrentProcessingTimeMs` simply returns the [batchProcessingTimeMs](#).

## Updating State — `update` Method

`update(newValue: S): Unit`

**Note**

`update` is part of the [GroupState Contract](#) to update the state.

`update ...FIXME`

## Removing State — `remove` Method

`remove(): Unit`

**Note**

`remove` is part of the [GroupState Contract](#) to remove the state.

`remove ...FIXME`

## Internal Properties

Name	Description
value	FIXME Used when...FIXME
defined	FIXME Used when...FIXME
updated	<b>Updated flag</b> that says whether the state has been <a href="#">updated</a> or not  Default: <code>false</code>  Disabled ( <code>false</code> ) when <code>GroupStateImpl</code> is requested to <a href="#">remove the state</a>  Enabled ( <code>true</code> ) when <code>GroupStateImpl</code> is requested to <a href="#">update the state</a>
removed	<b>Removed flag</b> that says whether the state is marked <a href="#">removed</a> or not  Default: <code>false</code>  Disabled ( <code>false</code> ) when <code>GroupStateImpl</code> is requested to <a href="#">update the state</a>  Enabled ( <code>true</code> ) when <code>GroupStateImpl</code> is requested to <a href="#">remove the state</a>
timeoutTimestamp	Current <b>timeout timestamp</b> (in millis) for <a href="#">GroupStateTimeout.EventTimeTimeout</a> or <a href="#">GroupStateTimeout.ProcessingTimeTimeout</a>  Default: <code>-1</code>  Defined using <a href="#">setTimeoutTimestamp</a> (for <code>EventTimeTimeout</code> ) and <a href="#">setTimeoutDuration</a> (for <code>ProcessingTimeTimeout</code> )

# GroupStateTimeout — Group State Timeout in Arbitrary Stateful Streaming Aggregation

`GroupStateTimeout` represents an **aggregation state timeout** that defines when a `GroupState` can be considered **timed-out (expired)** in `Arbitrary Stateful Streaming Aggregation`.

`GroupStateTimeout` is used with the following `KeyValueGroupedDataset` operations:

- `mapGroupsWithState`
- `flatMapGroupsWithState`

Table 1. GroupStateTimeouts

GroupStateTimeout	Description
<code>EventTimeTimeout</code>	Timeout based on event time Used when...FIXME
<code>NoTimeout</code>	No timeout Used when...FIXME
<code>ProcessingTimeTimeout</code>	Timeout based on processing time  <code>FlatMapGroupsWithStateExec</code> physical operator requires that <code>batchTimestampMs</code> is specified when <code>ProcessingTimeTimeout</code> is used.  <code>batchTimestampMs</code> is defined when <code>IncrementalExecution</code> is created (with the <code>state</code> ). <code>IncrementalExecution</code> is given <code>OffsetSeqMetadata</code> when <code>StreamExecution</code> is requested to run a streaming batch.

# StateManager Contract — State Managers for Arbitrary Stateful Streaming Aggregation

`StateManager` is the abstraction of state managers that act as *middlemen* between state stores and the `FlatMapGroupsWithStateExec` physical operator used in Arbitrary Stateful Streaming Aggregation.

Table 1. StateManager Contract

Method	Description
<code>getAllState</code>	<pre>getAllState(store: StateStore): Iterator[StateData]</pre> <p>Retrieves all state data (for all keys) from the <code>StateStore</code> Used exclusively when <code>InputProcessor</code> is requested to <code>processTimedOutState</code></p>
<code>getState</code>	<pre>getState(   store: StateStore,   keyRow: UnsafeRow): StateData</pre> <p>Gets the state data for the key from the <code>StateStore</code> Used exclusively when <code>InputProcessor</code> is requested to <code>processNewData</code></p>
<code>putState</code>	<pre>putState(   store: StateStore,   keyRow: UnsafeRow,   state: Any,   timeoutTimestamp: Long): Unit</pre> <p>Persists (<i>puts</i>) the state value for the key in the <code>StateStore</code> Used exclusively when <code>InputProcessor</code> is requested to <code>callFunctionAndUpdateState</code> (right after all rows have been processed)</p>
<code>removeState</code>	<pre>removeState(   store: StateStore,   keyRow: UnsafeRow): Unit</pre> <p>Removes the state for the key from the <code>StateStore</code></p>

	<p>Used exclusively when <code>InputProcessor</code> is requested to <code>callFunctionAndUpdateState</code> (right after all rows have been processed)</p>		
stateSchema	<p><code>stateSchema: StructType</code></p> <h3>State schema</h3> <table border="1"> <tr> <td>Note</td><td> <p>It looks like (in <code>StateManager</code> of the <code>FlatMapGroupsWithStateExec</code> physical operator) <code>stateSchema</code> is used for the schema of state value objects (not state keys as they are described by the grouping attributes instead).</p> </td></tr> </table> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>FlatMapGroupsWithStateExec</code> physical operator is requested to execute and generate a recipe for a distributed computation (as an <code>RDD[InternalRow]</code>)</li> <li>• <code>StateManagerImplBase</code> is requested for the <code>stateDeserializerFunc</code></li> </ul>	Note	<p>It looks like (in <code>StateManager</code> of the <code>FlatMapGroupsWithStateExec</code> physical operator) <code>stateSchema</code> is used for the schema of state value objects (not state keys as they are described by the grouping attributes instead).</p>
Note	<p>It looks like (in <code>StateManager</code> of the <code>FlatMapGroupsWithStateExec</code> physical operator) <code>stateSchema</code> is used for the schema of state value objects (not state keys as they are described by the grouping attributes instead).</p>		
Note	<p><code>StateManagerImplBase</code> is the one and only known direct implementation of the <code>StateManager Contract</code> in Spark Structured Streaming.</p>		
Note	<p><code>StateManager</code> is a Scala <b>sealed trait</b> which means that all the implementations are in the same compilation unit (a single file).</p>		

# StateManagerImplV2 — Default StateManager of FlatMapGroupsWithStateExec Physical Operator

`StateManagerImplV2` is a concrete `StateManager` (as a `StateManagerImplBase`) that is used by default in `FlatMapGroupsWithStateExec` physical operator (per `spark.sql.streaming.flatMapGroupsWithState.stateFormatVersion` internal configuration property).

`StateManagerImplV2` is `created` exclusively when `FlatMapGroupsWithStateExecHelper` utility is requested for a `StateManager` (when the `stateFormatVersion` is `2`).

## Creating StateManagerImplV2 Instance

`StateManagerImplV2` takes the following to be created:

- State encoder (`ExpressionEncoder[Any]`)
- `shouldStoreTimestamp` flag

`StateManagerImplV2` initializes the [internal properties](#).

## State Schema — stateSchema Value

```
stateSchema: StructType
```

Note	<code>stateSchema</code> is part of the <a href="#">StateManager Contract</a> for the schema of the state.
------	--

`stateSchema` ...FIXME

## State Serializer — stateSerializerExprs Value

```
stateSerializerExprs: Seq[Expression]
```

Note	<code>stateSerializerExprs</code> is part of the <a href="#">StateManager Contract</a> for the state serializer, i.e. Catalyst expressions to serialize a state object to a row ( <code>UnsafeRow</code> ).
------	---

`stateSerializerExprs` ...FIXME

## State Deserializer — `stateDeserializerExpr` Value

`stateDeserializerExpr`: Expression

Note	<code>stateDeserializerExpr</code> is part of the <a href="#">StateManager Contract</a> for the state deserializer, i.e. a Catalyst expression to deserialize a state object from a row ( <code>UnsafeRow</code> ).
------	---

`stateDeserializerExpr` ...FIXME

## Internal Properties

Name	Description
<code>nestedStateOrdinal</code>	Position of the state in a state row ( <code>0</code> ) Used when...FIXME
<code>timeoutTimestampOrdinalInRow</code>	Position of the timeout timestamp in a state row ( <code>1</code> ) Used when...FIXME

# StateManagerImplBase

`StateManagerImplBase` is the extension of the [StateManager contract](#) for [state managers](#) of [FlatMapGroupsWithStateExec](#) physical operator with the following features:

- Use Catalyst expressions for [state serialization](#) and [deserialization](#)
- Use `timeoutTimestampOrdinalInRow` when `shouldStoreTimestamp` with the `shouldStoreTimestamp` flag on

Table 1. StateManagerImplBase Contract (Abstract Methods Only)

Method	Description
<code>stateDeserializerExpr</code>	<code>stateDeserializerExpr: Expression</code> <b>State deserializer</b> , i.e. a Catalyst expression to deserialize a state object from a row ( <code>UnsafeRow</code> ) Used exclusively for the <a href="#">stateDeserializerFunc</a>
<code>stateSerializerExprs</code>	<code>stateSerializerExprs: Seq[Expression]</code> <b>State serializer</b> , i.e. Catalyst expressions to serialize a state object to a row ( <code>unsafeRow</code> ) Used exclusively for the <a href="#">stateSerializerFunc</a>
<code>timeoutTimestampOrdinalInRow</code>	<code>timeoutTimestampOrdinalInRow: Int</code> Position of the timeout timestamp in a state row Used when <code>StateManagerImplBase</code> is requested to <a href="#">get</a> and <a href="#">set timeout timestamp</a>

Table 2. StateManagerImplBases

StateManagerImplBase	Description
<a href="#">StateManagerImplV1</a>	Legacy <a href="#">StateManager</a>
<a href="#">StateManagerImplV2</a>	Default <a href="#">StateManager</a>

## Creating StateManagerImplBase Instance

`StateManagerImplBase` takes a single `shouldStoreTimestamp` flag to be created (that is set when the [concrete StateManagerImplBases](#) are created).

**Note**

`StateManagerImplBase` is a Scala abstract class and cannot be [created](#) directly. It is created indirectly for the [concrete StateManagerImplBases](#).

`StateManagerImplBase` initializes the [internal properties](#).

## Getting State Data for Key from StateStore — `getState` Method

```
getState(  
    store: StateStore,  
    keyRow: UnsafeRow): StateData
```

**Note**

`getState` is part of the [StateManager Contract](#) to get the state data for the key from the [StateStore](#).

`getState` ...FIXME

## Persisting State Value for Key in StateStore — `putState` Method

```
putState(  
    store: StateStore,  
    key: UnsafeRow,  
    state: Any,  
    timestamp: Long): Unit
```

**Note**

`putState` is part of the [StateManager Contract](#) to persist (*put*) the state value for the key in the [StateStore](#).

`putState` ...FIXME

## Removing State for Key from StateStore — `removeState` Method

```
removeState(  
    store: StateStore,  
    keyRow: UnsafeRow): Unit
```

**Note**

`removeState` is part of the [StateManager Contract](#) to remove the state for the key from the [StateStore](#).

`removeState` ...FIXME

## Getting All State Data (for All Keys) from StateStore

### — `getAllState` Method

```
getAllState(store: StateStore): Iterator[StateData]
```

**Note**

`getAllState` is part of the [StateManager Contract](#) to retrieve all state data (for all keys) from the [StateStore](#).

`getAllState` ...FIXME

## getStateObject Internal Method

```
getStateObject(row: UnsafeRow): Any
```

`getStateObject` ...FIXME

**Note**

`getStateObject` is used when...FIXME

## getStateRow Internal Method

```
getStateRow(obj: Any): UnsafeRow
```

`getStateRow` ...FIXME

**Note**

`getStateRow` is used when...FIXME

## Getting Timeout Timestamp (from State Row)

### — `getTimestamp` Internal Method

```
getTimestamp(stateRow: UnsafeRow): Long
```

`getTimestamp` ...FIXME

Note

`getTimestamp` is used when...FIXME

## Setting Timeout Timestamp (to State Row)

### — `setTimestamp` Internal Method

```
setTimestamp(
    stateRow: UnsafeRow,
    timeoutTimestamps: Long): Unit
```

`setTimestamp` ...FIXME

Note

`setTimestamp` is used when...FIXME

## Internal Properties

Name	Description
<code>stateSerializerFunc</code>	<p><b>State object serializer</b> (of type <code>Any ⇒ UnsafeRow</code>) to serialize a state object (for a per-group state key) to a row (<code>UnsafeRow</code>)</p> <ul style="list-style-type: none"> <li>The serialization expression (incl. the type) is specified as the <code>stateSerializerExprs</code></li> </ul> <p>Used exclusively in <a href="#">getStateRow</a></p>
<code>stateDeserializerFunc</code>	<p><b>State object deserializer</b> (of type <code>InternalRow ⇒ Any</code>) to deserialize a row (for a per-group state value) to a Scala value</p> <ul style="list-style-type: none"> <li>The deserialization expression (incl. the type) is specified as the <code>stateDeserializerExpr</code></li> </ul> <p>Used exclusively in <a href="#">getStateObject</a></p>
<code>stateDataForGets</code>	Empty <code>StateData</code> to share ( <i>reuse</i> ) between <a href="#">getState</a> calls (to avoid high use of memory with many <code>StateData</code> objects)

# StateManagerImplV1

StateManagerImplV1 is...FIXME

# FlatMapGroupsWithStateExecHelper

`FlatMapGroupsWithStateExecHelper` is a utility with the main purpose of creating a `StateManager` for `FlatMapGroupsWithStateExec` physical operator.

## Creating StateManager — `createStateManager` Method

```
createStateManager(  
    stateEncoder: ExpressionEncoder[Any],  
    shouldStoreTimestamp: Boolean,  
    stateFormatVersion: Int): StateManager
```

`createStateManager` simply creates a `StateManager` (with the `stateEncoder` and `shouldStoreTimestamp` flag) based on `stateFormatVersion`:

- `StateManagerImplV1` for 1
- `StateManagerImplV2` for 2

`createStateManager` throws an `IllegalArgumentException` for `stateFormatVersion` not 1 or 2:

```
Version [stateFormatVersion] is invalid
```

Note

`createStateManager` is used exclusively for the `StateManager` for `FlatMapGroupsWithStateExec` physical operator.

# InputProcessor Helper Class of FlatMapGroupsWithStateExec Physical Operator

`InputProcessor` is a helper class to manage state in the `state store` for every partition of a `FlatMapGroupsWithStateExec` physical operator.

`InputProcessor` is `created` exclusively when `FlatMapGroupsWithStateExec` physical operator is requested to `execute and generate a recipe for a distributed computation (as an RDD[InternalRow])` (and uses `InputProcessor` in the `storeUpdateFunction` while processing rows per partition with a corresponding per-partition state store).

`InputProcessor` takes a single `StateStore` to be created. The `statestore` manages the per-group state (and is used when processing `new data` and `timed-out state data`, and in the "all rows processed" callback).

## Processing New Data (Creating Iterator of New Data Processed) — `processNewData` Method

```
processNewData(dataIter: Iterator[InternalRow]): Iterator[InternalRow]
```

`processNewData` creates a grouped iterator of (of pairs of) per-group state keys and the row values from the given data iterator (`dataIter`) with the `grouping attributes` and the output schema of the `child operator` (of the parent `FlatMapGroupsWithStateExec` physical operator).

For every per-group state key (in the grouped iterator), `processNewData` requests the `StateManager` (of the parent `FlatMapGroupsWithStateExec` physical operator) to `get the state` (from the `StateStore`) and `callFunctionAndUpdateState` (with the `hasTimedOut` flag off, i.e. `false` ).

Note

`processNewData` is used exclusively when `FlatMapGroupsWithStateExec` physical operator is requested to `execute and generate a recipe for a distributed computation (as an RDD[InternalRow])`.

## Processing Timed-Out State Data (Creating Iterator of Timed-Out State Data) — `processTimedOutState` Method

```
processTimedOutState(): Iterator[InternalRow]
```

`processTimedOutState` does nothing and simply returns an empty iterator for `GroupStateTimeout.NoTimeout`.

With `timeout` enabled, `processTimedOutState` gets the current timeout threshold per `GroupStateTimeout`:

- `batchTimestampMs` for `ProcessingTimeTimeout`
- `eventTimeWatermark` for `EventTimeTimeout`

`processTimedOutState` creates an iterator of timed-out state data by requesting the `StateManager` for all the available state data (in the `StateStore`) and takes only the state data with timeout defined and below the current timeout threshold.

In the end, for every timed-out state data, `processTimedOutState` `callFunctionAndUpdateState` (with the `hasTimedOut` flag enabled).

Note	<code>processTimedOutState</code> is used exclusively when <code>FlatMapGroupsWithStateExec</code> physical operator is requested to execute and generate a recipe for a distributed computation (as an <code>RDD[InternalRow]</code> ).
------	--

## callFunctionAndUpdateState Internal Method

```
callFunctionAndUpdateState(
    stateData: StateData,
    valueRowIter: Iterator[InternalRow],
    hasTimedOut: Boolean): Iterator[InternalRow]
```

Note	<code>callFunctionAndUpdateState</code> is used when <code>InputProcessor</code> is requested to process new data and timed-out state data. When processing new data, <code>hasTimedOut</code> flag is off ( <code>false</code> ). When processing timed-out state data, <code>hasTimedOut</code> flag is on ( <code>true</code> ).
------	---

`callFunctionAndUpdateState` creates a key object by requesting the given `StateData` for the `UnsafeRow` of the key (`keyRow`) and converts it to an object (using the internal `state key converter`).

`callFunctionAndUpdateState` creates value objects by taking every value row (from the given `valueRowIter` iterator) and converts them to objects (using the internal `state value converter`).

`callFunctionAndUpdateState` creates a new `GroupStateImpl` with the following:

- The current state value (of the given `StateData`) that could possibly be `null`
- The `batchTimestampMs` of the parent `FlatMapGroupsWithStateExec` operator (that could possibly be `-1`)
- The `event-time watermark` of the parent `FlatMapGroupsWithStateExec` operator (that could possibly be `-1`)
- The `GroupStateTimeout` of the parent `FlatMapGroupsWithStateExec` operator
- The `watermarkPresent` flag of the parent `FlatMapGroupsWithStateExec` operator
- The given `hasTimedOut` flag

`callFunctionAndUpdateState` then executes the `user-defined state function` (of the parent `FlatMapGroupsWithStateExec` operator) on the key object, value objects, and the newly-created `GroupStateImpl`.

For every output value from the user-defined state function, `callFunctionAndUpdateState` updates `numOutputRows` performance metric and wraps the values to an internal row (using the internal `output value converter`).

In the end, `callFunctionAndUpdateState` returns a `Iterator[InternalRow]` which calls the `completion function` right after rows have been processed (so the iterator is considered fully consumed).

## "All Rows Processed" Callback — `onIteratorCompletion` Internal Method

```
onIteratorCompletion: Unit
```

`onIteratorCompletion` branches off per whether the `GroupStateImpl` has been marked `removed` and no `timeout timestamp` is specified or not.

When the `GroupStateImpl` has been marked `removed` and no `timeout timestamp` is specified, `onIteratorCompletion` does the following:

1. Requests the `StateManager` (of the parent `FlatMapGroupsWithStateExec` operator) to remove the state (from the `StateStore` for the key row of the given `StateData`)
2. Increments the `numUpdatedStateRows` performance metric

Otherwise, when the `GroupStateImpl` has not been marked `removed` or the `timeout timestamp` is specified, `onIteratorCompletion` checks whether the timeout timestamp has changed by comparing the timeout timestamps of the `GroupStateImpl` and the given `StateData`.

(only when the `GroupStateImpl` has been `updated`, `removed` or the timeout timestamp changed) `onIteratorCompletion` does the following:

1. Requests the `StateManager` (of the parent `FlatMapGroupsWithStateExec` operator) to `persist the state` (in the `StateStore` with the key row, updated state object, and the timeout timestamp of the given `StateData`)
2. Increments the `numUpdatedStateRows` performance metrics

Note	<code>onIteratorCompletion</code> is used exclusively when <code>InputProcessor</code> is requested to <code>callFunctionAndUpdateState</code> (right after rows have been processed)
------	---

## Internal Properties

Name	Description
getKeyObj	<p>A <b>state key converter</b> (of type <code>InternalRow = Any</code>) to deserialize a given row (for a per-group state key) to the current state value</p> <ul style="list-style-type: none"> <li>The deserialization expression for keys is specified as the <a href="#">key deserializer expression</a> when the parent <code>FlatMapGroupsWithStateExec</code> operator is created</li> <li>The data type of state keys is specified as the <a href="#">grouping attributes</a> when the parent <code>FlatMapGroupsWithStateExec</code> operator is created</li> </ul> <p>Used exclusively when <code>InputProcessor</code> is requested to <a href="#">callFunctionAndUpdateState</a>.</p>
getOutputRow	<p>A <b>output value converter</b> (of type <code>Any = InternalRow</code>) to wrap a given output value (from the user-defined state function) to a row</p> <ul style="list-style-type: none"> <li>The data type of the row is specified as the data type of the <a href="#">output object attribute</a> when the parent <code>FlatMapGroupsWithStateExec</code> operator is created</li> </ul> <p>Used exclusively when <code>InputProcessor</code> is requested to <a href="#">callFunctionAndUpdateState</a>.</p>
getValueObj	<p>A <b>state value converter</b> (of type <code>InternalRow = Any</code>) to deserialize a given row (for a per-group state value) to a Scala value</p> <ul style="list-style-type: none"> <li>The deserialization expression for values is specified as the <a href="#">value deserializer expression</a> when the parent <code>FlatMapGroupsWithStateExec</code> operator is created</li> <li>The data type of state values is specified as the <a href="#">data attributes</a> when the parent <code>FlatMapGroupsWithStateExec</code> operator is created</li> </ul> <p>Used exclusively when <code>InputProcessor</code> is requested to <a href="#">callFunctionAndUpdateState</a>.</p>
numOutputRows	<code>numOutputRows</code> performance metric

# DataStreamReader — Loading Data from Streaming Source

`DataStreamReader` is the [interface](#) to describe how data is [loaded](#) to a streaming `Dataset` from a [streaming source](#).

Table 1. DataStreamReader's Methods

Method	Description
<code>csv</code>	<pre>csv(path: String): DataFrame</pre> <p>Sets <code>csv</code> as the <a href="#">format</a> of the data source</p>
<code>format</code>	<pre>format(source: String): DataStreamReader</pre> <p>Specifies the format of the <a href="#">data source</a> The format is used internally as the name (<i>alias</i>) of the <a href="#">streaming source</a> to use to load the data</p>
<code>json</code>	<pre>json(path: String): DataFrame</pre> <p>Sets <code>json</code> as the <a href="#">format</a> of the data source</p>
<code>load</code>	<pre>load(): DataFrame load(path: String): DataFrame (1)</pre> <p>1. Explicit <code>path</code> (that could also be specified as an <a href="#">option</a>) Creates a streaming <code>DataFrame</code> that represents "loading" streaming data (and is internally a logical plan with a <a href="#">StreamingRelationV2</a> or <a href="#">StreamingRelation</a> leaf logical operators)</p>
<code>option</code>	<pre>option(key: String, value: Boolean): DataStreamReader option(key: String, value: Double): DataStreamReader option(key: String, value: Long): DataStreamReader option(key: String, value: String): DataStreamReader</pre> <p>Sets a loading option</p>
	<pre>options(options: Map[String, String]): DataStreamReader</pre>

	<code>options</code>	Specifies the configuration options of a data source		
		<table border="1"> <tr> <td>Note</td> <td>You could use <code>option</code> method if you prefer specifying the options one by one or there is only one in use.</td> </tr> </table>	Note	You could use <code>option</code> method if you prefer specifying the options one by one or there is only one in use.
Note	You could use <code>option</code> method if you prefer specifying the options one by one or there is only one in use.			
	<code>orc</code>	<pre>orc(path: String): DataFrame</pre> <p>Sets <code>orc</code> as the <code>format</code> of the data source</p>		
	<code>parquet</code>	<pre>parquet(path: String): DataFrame</pre> <p>Sets <code>parquet</code> as the <code>format</code> of the data source</p>		
	<code>schema</code>	<pre>schema(schema: StructType): DataStreamReader schema(schemaString: String): DataStreamReader (1)</pre> <p>1. Uses a DDL-formatted table schema</p> <p>Specifies the <code>user-defined schema</code> of the streaming data source (as a <code>StructType</code> or DDL-formatted table schema, e.g. <code>a INT, b STRING</code>)</p>		
	<code>text</code>	<pre>text(path: String): DataFrame</pre> <p>Sets <code>text</code> as the <code>format</code> of the data source</p>		
	<code>textFile</code>	<pre>textFile(path: String): Dataset[String]</pre>		



Figure 1. DataStreamReader and The Others

`DataStreamReader` is used for a Spark developer to describe how Spark Structured Streaming loads datasets from a streaming source (that `in the end` creates a logical plan for a streaming query).

Note	<code>DataStreamReader</code> is the Spark developer-friendly API to create a <code>StreamingRelation</code> logical operator (that represents a <code>streaming source</code> in a logical plan).
------	--

You can access `DataStreamReader` using `SparkSession.readStream` method.

```
import org.apache.spark.sql.SparkSession
val spark: SparkSession = ...

val streamReader = spark.readStream
```

`DataStreamReader` supports many [source formats](#) natively and offers the [interface to define custom formats](#):

- [json](#)
- [csv](#)
- [parquet](#)
- [text](#)

Note	<code>DataStreamReader</code> assumes <a href="#">parquet</a> file format by default that you can change using <code>spark.sql.sources.default</code> property.
------	---

Note	<code>hive</code> source format is not supported.
------	---

After you have described the **streaming pipeline** to read datasets from an external streaming data source, you eventually trigger the loading using format-agnostic [load](#) or format-specific (e.g. [json](#), [csv](#)) operators.

Table 2. DataStreamReader's Internal Properties (in alphabetical order)

Name	Initial Value	Description
<code>source</code>	<code>spark.sql.sources.default</code> property	Source format of datasets in a streaming data source
<code>userSpecifiedSchema</code>	(empty)	Optional user-defined schema
<code>extraOptions</code>	(empty)	Collection of key-value configuration options

## Specifying Loading Options — `option` Method

```
option(key: String, value: String): DataStreamReader
option(key: String, value: Boolean): DataStreamReader
option(key: String, value: Long): DataStreamReader
option(key: String, value: Double): DataStreamReader
```

`option` family of methods specifies additional options to a streaming data source.

There is support for values of `String`, `Boolean`, `Long`, and `Double` types for user convenience, and internally are converted to `String` type.

Internally, `option` sets [extraOptions](#) internal property.

**Note**

You can also set options in bulk using [options](#) method. You have to do the type conversion yourself, though.

## Creating Streaming Dataset (to Represent Loading Data From Streaming Source) — `load` Method

```
load(): DataFrame
load(path: String): DataFrame (1)
```

1. Specifies `path` option before passing the call to parameterless `load()`

```
load ...FIXME
```

## Built-in Formats

```
json(path: String): DataFrame
csv(path: String): DataFrame
parquet(path: String): DataFrame
text(path: String): DataFrame
textFile(path: String): Dataset[String] (1)
```

1. Returns `Dataset[String]` not `DataFrame`

`DataStreamReader` can load streaming datasets from data sources of the following [formats](#):

- `json`
- `csv`
- `parquet`
- `text`

The methods simply pass calls to [format](#) followed by `load(path)`.

# DataStreamWriter — Writing Datasets To Streaming Sink

`DataStreamWriter` is the [interface](#) to describe when and what rows of a streaming query are sent out to the [streaming sink](#).

`DataStreamWriter` is available using [Dataset.writeStream](#) method (on a streaming query).

```
import org.apache.spark.sql.streaming.DataStreamWriter
import org.apache.spark.sql.Row

val streamingQuery: Dataset[Long] = ...

assert(streamingQuery.isStreaming)

val writer: DataStreamWriter[Row] = streamingQuery.writeStream
```

Table 1. DataStreamWriter's Methods

Method	Description
<code>foreach</code>	<pre>foreach(writer: ForeachWriter[T]): DataStreamWriter[T]</pre> <p>Sets <a href="#">ForeachWriter</a> in the full control of streaming writes</p>
<code>foreachBatch</code>	<pre>foreachBatch(     function: (Dataset[T], Long) =&gt; Unit): DataStreamWriter[T]</pre> <p><b>(New in 2.4.0)</b> Sets the <a href="#">source</a> to <code>foreachBatch</code> and the <a href="#">foreachBatchWriter</a> to the given function.</p> <p>As per <a href="#">SPARK-24565 Add API for in Structured Streaming for exposing output rows of each microbatch as a DataFrame</a>, the purpose of the method is to expose the micro-batch output as a dataframe for the following:</p> <ul style="list-style-type: none"> <li>• Pass the output rows of each batch to a library that is designed for the batch jobs only</li> <li>• Reuse batch data sources for output whose streaming version does not exist</li> <li>• Multi-writes where the output rows are written to multiple outputs by writing twice for every batch</li> </ul>
	<pre>format(source: String): DataStreamWriter[T]</pre>

<code>format</code>	<p>Specifies the format of the <a href="#">data sink</a> (aka <i>output format</i>)</p> <p>The format is used internally as the name (<i>alias</i>) of the <a href="#">streaming sink</a> to use to write the data to</p>		
<code>option</code>	<pre>option(key: String, value: Boolean): DataStreamWriter[T] option(key: String, value: Double): DataStreamWriter[T] option(key: String, value: Long): DataStreamWriter[T] option(key: String, value: String): DataStreamWriter[T]</pre>		
<code>options</code>	<pre>options(options: Map[String, String]): DataStreamWriter[T]</pre> <p>Specifies the configuration options of a data sink</p> <table border="1"> <tr> <td><b>Note</b></td><td>You could use <a href="#">option</a> method if you prefer specifying the options one by one or there is only one in use.</td></tr> </table>	<b>Note</b>	You could use <a href="#">option</a> method if you prefer specifying the options one by one or there is only one in use.
<b>Note</b>	You could use <a href="#">option</a> method if you prefer specifying the options one by one or there is only one in use.		
<code>outputMode</code>	<pre>outputMode(outputMode: OutputMode): DataStreamWriter[T] outputMode(outputMode: String): DataStreamWriter[T]</pre> <p>Specifies the <a href="#">output mode</a></p>		
<code>partitionBy</code>	<pre>partitionBy(colNames: String*): DataStreamWriter[T]</pre>		
<code>queryName</code>	<pre>queryName(queryName: String): DataStreamWriter[T]</pre> <p>Assigns the name of a query</p>		
<code>start</code>	<pre>start(): StreamingQuery start(path: String): StreamingQuery (1)</pre> <ol style="list-style-type: none"> <li>1. Explicit <code>path</code> (that could also be specified as an <a href="#">option</a>)</li> </ol> <p>Creates and immediately starts a <a href="#">StreamingQuery</a></p>		
<code>trigger</code>	<pre>trigger(trigger: Trigger): DataStreamWriter[T]</pre> <p>Sets the <a href="#">Trigger</a> for how often a streaming query should be executed and the result saved.</p>		

**Note**

A streaming query is a `Dataset` with a [streaming logical plan](#).

```

import org.apache.spark.sql.streaming.Trigger
import scala.concurrent.duration._
import org.apache.spark.sql.DataFrame
val rates: DataFrame = spark.
  readStream.
  format("rate").
  load

scala> rates.isStreaming
res1: Boolean = true

scala> rates.queryExecution.logical.isStreaming
res2: Boolean = true

```

Like the batch `DataFrameWriter`, `DataStreamWriter` has a direct support for many [file formats](#) and an extension point to plug in new formats.

```

// see above for writer definition

// Save dataset in JSON format
writer.format("json")

```

In the end, you start the actual continuous writing of the result of executing a `dataset` to a sink using `start` operator.

```
writer.save
```

Beside the above operators, there are the following to work with a `dataset` as a whole.

**Note**

`hive` is not supported for streaming writing (and leads to a `AnalysisException`).

**Note**

`DataStreamWriter` is responsible for writing in a batch fashion.

## Specifying Write Option — `option` Method

```

option(key: String, value: String): DataStreamWriter[T]
option(key: String, value: Boolean): DataStreamWriter[T]
option(key: String, value: Long): DataStreamWriter[T]
option(key: String, value: Double): DataStreamWriter[T]

```

Internally, `option` adds the `key` and `value` to `extraOptions` internal option registry.

## Specifying Output Mode — `outputMode` Method

```
outputMode(outputMode: String): DataStreamWriter[T]
outputMode(outputMode: OutputMode): DataStreamWriter[T]
```

`outputMode` specifies the [output mode](#) of a streaming query, i.e. what data is sent out to a [streaming sink](#) when there is new data available in [streaming data sources](#).

Note	When not defined explicitly, <code>outputMode</code> defaults to <a href="#">Append</a> output mode.
------	--

`outputMode` can be specified by name or one of the [OutputMode](#) values.

## Setting Query Name — `queryName` method

```
queryName(queryName: String): DataStreamWriter[T]
```

`queryName` sets the name of a [streaming query](#).

Internally, it is just an additional [option](#) with the key `queryName`.

## Setting How Often to Execute Streaming Query — `trigger` method

```
trigger(trigger: Trigger): DataStreamWriter[T]
```

`trigger` method sets the time interval of the [trigger](#) (that executes a batch runner) for a streaming query.

Note	<code>Trigger</code> specifies how often results should be produced by a <a href="#">StreamingQuery</a> . See <a href="#">Trigger</a> .
------	---

The default trigger is [ProcessingTime\(0L\)](#) that runs a streaming query as often as possible.

Tip	Consult <a href="#">Trigger</a> to learn about <code>Trigger</code> and <code>ProcessingTime</code> types.
-----	--

## Creating and Starting Execution of Streaming Query — `start` Method

```
start(): StreamingQuery
start(path: String): StreamingQuery (1)
```

1. Sets `path` option to `path` and passes the call on to `start()`

`start` starts a streaming query.

`start` gives a [StreamingQuery](#) to control the execution of the continuous query.

Note	Whether or not you have to specify <code>path</code> option depends on the streaming sink in use.
------	---

Internally, `start` branches off per `source`.

- `memory`
- `foreach`
- other formats

...FIXME

Table 2. `start`'s Options

Option	Description
<code>queryName</code>	Name of active streaming query
<code>checkpointLocation</code>	Directory for checkpointing (and to store query metadata like offsets before and after being processed, the <a href="#">query id</a> , etc.)

`start` reports a [AnalysisException](#) when `source` is `hive`.

```
val q = spark.  
  readStream.  
  text("server-logs/*").  
  writeStream.  
  format("hive") <-- hive format used as a streaming sink  
scala> q.start  
org.apache.spark.sql.AnalysisException: Hive data source can only be used with tables,  
you can not write files of Hive data source directly.;  
  at org.apache.spark.sql.streaming.DataStreamWriter.start(DataStreamWriter.scala:234)  
  ... 48 elided
```

Note	Define options using <code>option</code> or <code>options</code> methods.
------	---

## Making ForeachWriter in Charge of Streaming Writes

### — `foreach` method

```
foreach(writer: ForeachWriter[T]): DataStreamWriter[T]
```

`foreach` sets the input [ForeachWriter](#) to be in control of streaming writes.

Internally, `foreach` sets the streaming output [format](#) as `foreach` and `foreachWriter` as the input `writer`.

**Note**

`foreach` uses `SparkSession` to access `SparkContext` to clean the [ForeachWriter](#).

**Note**

`foreach` reports an `IllegalArgumentException` when `writer` is `null`.

`foreach writer` cannot be `null`

## Internal Properties

Name	Initial Value	Description
<code>extraOptions</code>		
<code>foreachBatchWriter</code>	<code>null</code>	<code>foreachBatchWriter: (Dataset[T], Long) =&gt; Unit</code> The function that is used as the batch writer in the <a href="#">ForeachBatchSink</a> for <code>foreachBatch</code>
<code>foreachWriter</code>		
<code>partitioningColumns</code>		
<code>source</code>		
<code>outputMode</code>	<a href="#">Append</a>	<a href="#">OutputMode</a> of the streaming sink Set using <a href="#">outputMode</a> method.
<code>trigger</code>		

# OutputMode

**Output mode** (`OutputMode`) of a streaming query describes what data is written to a [streaming sink](#).

There are three available output modes:

- [Append](#)
- [Complete](#)
- [Update](#)

The output mode is specified on the *writing side* of a streaming query using [DataStreamWriter.outputMode](#) method (by alias or a value of `org.apache.spark.sql.streaming.OutputMode` object).

```
import org.apache.spark.sql.streaming.OutputMode.Update
val inputStream = spark
  .readStream
  .format("rate")
  .load
  .writeStream
  .format("console")
  .outputMode(Update) // <-- update output mode
  .start
```

## Append Output Mode

**Append** (alias: `append`) is the [default output mode](#) that writes "new" rows only.

In [streaming aggregations](#), a "new" row is when the intermediate state becomes final, i.e. when new events for the grouping key can only be considered late which is when watermark moves past the event time of the key.

`Append` output mode requires that a streaming query defines event-time watermark (using [withWatermark](#) operator) on the event time column that is used in aggregation (directly or using [window](#) function).

Required for datasets with `FileFormat` format (to create [FileStreamSink](#))

`Append` is [mandatory](#) when multiple `flatMapGroupsWithState` operators are used in a structured query.

## Complete Output Mode

**Complete** (alias: **complete**) writes all the rows of a Result Table (and corresponds to a traditional batch structured query).

Complete mode does not drop old aggregation state and preserves all data in the Result Table.

Supported only for [streaming aggregations](#) (as asserted by [UnsupportedOperationChecker](#)).

## Update Output Mode

**Update** (alias: **update**) writes only the rows that were updated (every time there are updates).

For queries that are not [streaming aggregations](#), `Update` is equivalent to the [Append](#) output mode.

# Trigger — How Frequently to Check Sources For New Data

`Trigger` defines how often a `streaming query` should be executed (*triggered*) and emit a new data (which `StreamExecution` uses to `resolve a TriggerExecutor`).

Table 1. Trigger's Factory Methods

Trigger	Creating Instance
<code>ContinuousTrigger</code>	<code>Trigger Continuous(long intervalMs)</code> <code>Trigger Continuous(long interval, TimeUnit timeUnit)</code> <code>Trigger Continuous(Duration interval)</code> <code>Trigger Continuous(String interval)</code>
<code>OneTimeTrigger</code>	<code>Trigger Once()</code>
<code>ProcessingTime</code>	<code>Trigger ProcessingTime(Duration interval)</code> <code>Trigger ProcessingTime(long intervalMs)</code> <code>Trigger ProcessingTime(long interval, TimeUnit timeUnit)</code> <code>Trigger ProcessingTime(String interval)</code>
	<a href="#">Examples of ProcessingTime</a>
Note	You specify the trigger for a streaming query using <code>DataStreamWriter</code> 's <code>trigger</code> method.

```

import org.apache.spark.sql.streaming.Trigger
val query = spark.
  readStream.
  format("rate").
  load.
  writeStream.
  format("console").
  option("truncate", false).
  trigger(Trigger.Once). // <-- execute once and stop
  queryName("rate-once").
  start

assert(query.isActive == false)

scala> println(query.lastProgress)
{
  "id" : "2ae4b0a4-434f-4ca7-a523-4e859c07175b",
  "runId" : "24039ce5-906c-4f90-b6e7-bbb3ec38a1f5",
  "name" : "rate-once",
  "timestamp" : "2017-07-04T18:39:35.998Z",
  "numInputRows" : 0,
  "processedRowsPerSecond" : 0.0,
  "durationMs" : {
    "addBatch" : 1365,
    "getBatch" : 29,
    "getOffset" : 0,
    "queryPlanning" : 285,
    "triggerExecution" : 1742,
    "walCommit" : 40
  },
  "stateOperators" : [ ],
  "sources" : [ {
    "description" : "RateSource[rowsPerSecond=1, rampUpTimeSeconds=0, numPartitions=8]"
  ,
    "startOffset" : null,
    "endOffset" : 0,
    "numInputRows" : 0,
    "processedRowsPerSecond" : 0.0
  } ],
  "sink" : {
    "description" : "org.apache.spark.sql.execution.streaming.ConsoleSink@7dbf277"
  }
}

```

**Note**

Although `Trigger` allows for custom implementations, `StreamExecution` refuses such attempts and reports an `IllegalStateException`.

```

import org.apache.spark.sql.streaming.Trigger
case object MyTrigger extends Trigger
scala> val sq = spark
    .readStream
    .format("rate")
    .load
    .writeStream
    .format("console")
    .trigger(MyTrigger) // <-- use custom trigger
    .queryName("rate-custom-trigger")
    .start
java.lang.IllegalStateException: Unknown type of trigger: MyTrigger
  at org.apache.spark.sql.execution.streaming.MicroBatchExecution.<init>(MicroBatchExecution.scala:60)
  at org.apache.spark.sql.streaming.StreamingQueryManager.createQuery(StreamingQueryManager.scala:275)
  at org.apache.spark.sql.streaming.StreamingQueryManager.startQuery(StreamingQueryManager.scala:316)
  at org.apache.spark.sql.streaming.DataStreamWriter.start(DataStreamWriter.scala:325)
... 57 elided

```

**Note**

`Trigger` was introduced in [the commit for \[SPARK-14176\]\[SQL\] Add DataFrameWriter.trigger to set the stream batch period.](#)

## Examples of ProcessingTime

`ProcessingTime` is a `Trigger` that assumes that milliseconds is the minimum time unit.

You can create an instance of `ProcessingTime` using the following constructors:

- `ProcessingTime(Long)` that accepts non-negative values that represent milliseconds.

```
ProcessingTime(10)
```

- `ProcessingTime(interval: String)` or `ProcessingTime.create(interval: String)` that accept `CalendarInterval` instances with or without leading `interval` string.

```
ProcessingTime("10 milliseconds")
ProcessingTime("interval 10 milliseconds")
```

- `ProcessingTime(Duration)` that accepts `scala.concurrent.duration.Duration` instances.

```
ProcessingTime(10.seconds)
```

- `ProcessingTime.create(interval: Long, unit: TimeUnit)` for `Long` and `java.util.concurrent.TimeUnit` instances.

```
ProcessingTime.create(10, TimeUnit.SECONDS)
```

# StreamingQuery Contract

`StreamingQuery` is the [contract](#) of streaming queries that are executed continuously and concurrently (i.e. on a [separate thread](#)).

Note	<code>StreamingQuery</code> is called <b>continuous query</b> or <b>streaming query</b> .
------	---

Note	<code>StreamingQuery</code> is a Scala trait with the only implementation being <a href="#">StreamExecution</a> (and less importantly <a href="#">StreamingQueryWrapper</a> for serializing a non-serializable <code>streamExecution</code> ).
------	--

Table 1. StreamingQuery Contract

Method	Description
<code>awaitTermination</code>	<pre>awaitTermination(): Unit awaitTermination(timeoutMs: Long): Boolean</pre> <p>Used when...FIXME</p>
<code>exception</code>	<pre>exception: Option[StreamingQueryException]</pre> <p><code>StreamingQueryException</code> if the query has finished due to an exception</p> <p>Used when...FIXME</p>
<code>explain</code>	<pre>explain(): Unit explain(extended: Boolean): Unit</pre> <p>Used when...FIXME</p>
<code>id</code>	<pre>id: UUID</pre> <p>The unique identifier of the streaming query (that does not change across restarts unlike <a href="#">runId</a>)</p> <p>Used when...FIXME</p>
<code>isActive</code>	<pre>isActive: Boolean</pre> <p>Indicates whether the streaming query is active (<code>true</code>) or not (<code>false</code>)</p>

	Used when...FIXME
lastProgress	<pre>lastProgress: StreamingQueryProgress</pre> <p>The last <a href="#">StreamingQueryProgress</a> of the streaming query</p>
	Used when...FIXME
name	<pre>name: String</pre> <p>The name of the query that is unique across all active queries</p>
	Used when...FIXME
processAllAvailable	<pre>processAllAvailable(): Unit</pre> <p>Pauses (<i>blocks</i>) the current thread until the streaming query has no more data to be processed or has been <a href="#">stopped</a></p> <p>Intended for testing</p>
recentProgress	<pre>recentProgress: Array[StreamingQueryProgress]</pre> <p>Collection of the recent <a href="#">StreamingQueryProgress</a> updates.</p>
	Used when...FIXME
runId	<pre>runId: UUID</pre> <p>The unique identifier of the current execution of the streaming query (that is different every restart unlike <a href="#">id</a>)</p>
	Used when...FIXME
sparkSession	<pre>sparkSession: SparkSession</pre>
	Used when...FIXME
	<pre>status: StreamingQueryStatus</pre>

<code>status</code>	<code>StreamingQueryStatus</code> of the streaming query (as <code>StreamExecution</code> <code>has accumulated</code> being a <code>ProgressReporter</code> while running the streaming query)  Used when...FIXME
<code>stop</code>	<code>stop(): Unit</code>  Stops the streaming query

`StreamingQuery` can be in two states:

- active (started)
- inactive (stopped)

If inactive, `StreamingQuery` may have transitioned into the state due to an `StreamingQueryException` (that is available under `exception` ).

`StreamingQuery` tracks current state of all the sources, i.e. `SourceStatus`, as `sourceStatuses`.

There could only be a single `Sink` for a `StreamingQuery` with many `Sources`.

`StreamingQuery` can be stopped by `stop` or an exception.

# Streaming Operators — High-Level Declarative Streaming Dataset API

Dataset API comes with a set of [operators](#) that are of particular use in Spark Structured Streaming that together constitute so-called **High-Level Declarative Streaming Dataset API**.

Table 1. Streaming Operators

Operator	Description
<code>crossJoin</code>	<code>crossJoin(     right: Dataset[_]): DataFrame</code>
<code>dropDuplicates</code>	<code>dropDuplicates(): Dataset[T] dropDuplicates(colNames: Seq[String]): Dataset[T] dropDuplicates(col1: String, cols: String*): Dataset[T]</code>  Drops duplicate records (given a subset of columns)
<code>explain</code>	<code>explain(): Unit explain(extended: Boolean): Unit</code>  Explains query plans
<code>groupBy</code>	<code>groupBy(cols: Column*): RelationalGroupedDataset groupBy(col1: String, cols: String*): RelationalGroupedDataset</code>  Aggregates rows by zero, one or more columns
<code>groupByKey</code>	<code>groupByKey(func: T =&gt; K): KeyValueGroupedDataset[K, T]</code>  Aggregates rows by a typed grouping function (and gives a <a href="#">KeyValueGroupedDataset</a> )

	<pre> join(     right: Dataset[_]): DataFrame join(     right: Dataset[_],     joinExprs: Column): DataFrame join(     right: Dataset[_],     joinExprs: Column,     joinType: String): DataFrame join(     right: Dataset[_],     usingColumns: Seq[String]): DataFrame join(     right: Dataset[_],     usingColumns: Seq[String],     joinType: String): DataFrame join(     right: Dataset[_],     usingColumn: String): DataFrame </pre>
joinWith	<pre> joinWith[U](     other: Dataset[U],     condition: Column): Dataset[(T, U)] joinWith[U](     other: Dataset[U],     condition: Column,     joinType: String): Dataset[(T, U)] </pre>
withWatermark	<pre> withWatermark(     eventTime: String,     delayThreshold: String): Dataset[T] </pre> <p>Defines a <a href="#">streaming watermark</a> (on the given <code>eventTime</code> column with a delay threshold)</p>
writeStream	<pre> writeStream: DataStreamWriter[T] </pre> <p>Creates a <a href="#">DataStreamWriter</a> for persisting the result of a streaming query to an external data system</p>

```
val rates = spark
    .readStream
    .format("rate")
    .option("rowsPerSecond", 1)
    .load

// stream processing
// replace [operator] with the operator of your choice
rates.[operator]

// output stream
import org.apache.spark.sql.streaming.{OutputMode, Trigger}
import scala.concurrent.duration._
val sq = rates
    .writeStream
    .format("console")
    .option("truncate", false)
    .trigger(Trigger.ProcessingTime(10.seconds))
    .outputMode(OutputMode.Complete)
    .queryName("rate-console")
    .start

// eventually...
sq.stop
```

# dropDuplicates Operator — Streaming Deduplication

```
dropDuplicates(): Dataset[T]
dropDuplicates(colNames: Seq[String]): Dataset[T]
dropDuplicates(col1: String, cols: String*): Dataset[T]
```

`dropDuplicates` operator...FIXME

Note	For a streaming Dataset, <code>dropDuplicates</code> will keep all data across triggers as intermediate state to drop duplicates rows. You can use <a href="#">withWatermark</a> operator to limit how late the duplicate data can be and system will accordingly limit the state. In addition, too late data older than watermark will be dropped to avoid any possibility of duplicates.
------	--

```
scala> spark.version
res0: String = 2.3.0-SNAPSHOT

// Start a streaming query
// Using old-fashioned MemoryStream (with the deprecated SQLContext)
import org.apache.spark.sql.execution.streaming.MemoryStream
import org.apache.spark.sql.SQLContext
implicit val sqlContext: SQLContext = spark.sqlContext
val source = MemoryStream[(Int, Int)]
val ids = source.toDS.toDF("time", "id").
  withColumn("time", $"time" cast "timestamp"). // <-- convert time column from Int to
  Timestamp
  dropDuplicates("id").
  withColumn("time", $"time" cast "long") // <-- convert time column back from Timest
amp to Int

// Conversions are only for display purposes
// Internally we need timestamps for watermark to work
// Displaying timestamps could be too much for such a simple task

scala> println(ids.queryExecution.analyzed.numberedTreeString)
00 Project [cast(time#10 as bigint) AS time#15L, id#6]
01 +- Deduplicate [id#6], true
02   +- Project [cast(timestamp#5 as timestamp) AS time#10, id#6]
03     +- Project [_1#2 AS time#5, _2#3 AS id#6]
04       +- StreamingExecutionRelation MemoryStream[_1#2,_2#3], [_1#2, _2#3]

import org.apache.spark.sql.streaming.{OutputMode, Trigger}
import scala.concurrent.duration.-
val q = ids.
  writeStream.
    format("memory").
```

```
queryName("dups").
  outputMode(OutputMode.Append).
  trigger(Trigger.ProcessingTime(30.seconds)).
  option("checkpointLocation", "checkpoint-dir"). // <-- use checkpointing to save state between restarts
  start

// Publish duplicate records
source.addData(1 -> 1)
source.addData(2 -> 1)
source.addData(3 -> 1)

q.processAllAvailable()

// Check out how dropDuplicates removes duplicates
// --> per single streaming batch (easy)
scala> spark.table("dups").show
+---+---+
|time| id|
+---+---+
|  1|  1|
+---+---+

source.addData(4 -> 1)
source.addData(5 -> 2)

// --> across streaming batches (harder)
scala> spark.table("dups").show
+---+---+
|time| id|
+---+---+
|  1|  1|
|  5|  2|
+---+---+

// Check out the internal state
scala> println(q.lastProgress.stateOperators(0).prettyJson)
{
  "numRowsTotal" : 2,
  "numRowsUpdated" : 1,
  "memoryUsedBytes" : 17751
}

// You could use web UI's SQL tab instead
// Use Details for Query

source.addData(6 -> 2)

scala> spark.table("dups").show
+---+---+
|time| id|
+---+---+
|  1|  1|
```

```

|   5|  2|
+----+---+
// Check out the internal state
scala> println(q.lastProgress.stateOperators(0).prettyJson)
{
  "numRowsTotal" : 2,
  "numRowsUpdated" : 0,
  "memoryUsedBytes" : 17751
}

// Restart the streaming query
q.stop

val q = ids.
  writeStream.
  format("memory").
  queryName("dups").
  outputMode(OutputMode.Complete). // <-- memory sink supports checkpointing for Complete output mode only
  trigger(Trigger.ProcessingTime(30.seconds)).
  option("checkpointLocation", "checkpoint-dir"). // <-- use checkpointing to save state between restarts
  start

// Doh! MemorySink is fine, but Complete is only available with a streaming aggregation

// Answer it if you know why --> https://stackoverflow.com/q/45756997/1305344

// It's a high time to work on https://issues.apache.org/jira/browse/SPARK-21667
// to understand the low-level details (and the reason, it seems)

// Disabling operation checks and starting over
// ./bin/spark-shell -c spark.sql.streaming.unsupportedOperationCheck=false
// it works now --> no exception!

scala> spark.table("dups").show
+----+---+
|time| id|
+----+---+
+----+---+

source.addData(0 -> 1)
// wait till the batch is triggered
scala> spark.table("dups").show
+----+---+
|time| id|
+----+---+
|   0|  1|
+----+---+

source.addData(1 -> 1)
source.addData(2 -> 1)

```

## dropDuplicates Operator

---

```
// wait till the batch is triggered
scala> spark.table("dups").show
+---+---+
|time| id|
+---+---+
+---+---+

// What?! No rows?! It doesn't look as if it worked fine :(

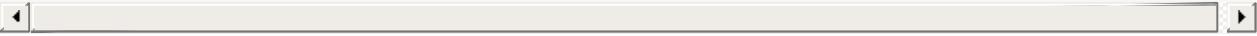
// Use groupBy to pass the requirement of having streaming aggregation for Complete ou
tput mode
val counts = ids.groupBy("id").agg(first($"time") as "first_time")
scala> counts.explain
== Physical Plan ==
*HashAggregate(keys=[id#246], functions=[first(time#255L, false)])
+- StateStoreSave [id#246], StatefulOperatorStateInfo(<unknown>, 3585583b-42d7-4547-8d62
-255581c48275, 0, 0), Append, 0
    +- *HashAggregate(keys=[id#246], functions=[merge_first(time#255L, false)])
        +- StateStoreRestore [id#246], StatefulOperatorStateInfo(<unknown>, 3585583b-42d7
-4547-8d62-255581c48275, 0, 0)
            +- *HashAggregate(keys=[id#246], functions=[merge_first(time#255L, false)])
                +- *HashAggregate(keys=[id#246], functions=[partial_first(time#255L, false
)])
                    +- *Project [cast(time#250 as bigint) AS time#255L, id#246]
                        +- StreamingDeduplicate [id#246], StatefulOperatorStateInfo(<unknow
n>, 3585583b-42d7-4547-8d62-255581c48275, 1, 0), 0
                            +- Exchange hashpartitioning(id#246, 200)
                                +- *Project [cast(_1#242 as timestamp) AS time#250, _2#243 AS
id#246]
                                    +- StreamingRelation MemoryStream[_1#242,_2#243], [_1#242,
_2#243]
val q = counts.
  writeStream.
format("memory").
queryName("dups").
outputMode(OutputMode.Complete). // <-- memory sink supports checkpointing for Comp
lete output mode only
  trigger(Trigger.ProcessingTime(30.seconds)).
  option("checkpointLocation", "checkpoint-dir"). // <-- use checkpointing to save sta
te between restarts
  start

source.addData(0 -> 1)
source.addData(1 -> 1)
// wait till the batch is triggered
scala> spark.table("dups").show
+---+-----+
| id|first_time|
+---+-----+
|  1|         0|
+---+-----+
// Publish duplicates
```

## dropDuplicates Operator

---

```
// Check out how dropDuplicates removes duplicates  
  
// Stop the streaming query  
// Specify event time watermark to remove old duplicates
```



# Dataset.explain High-Level Operator — Explaining Streaming Query Plans

```
explain(): Unit (1)
explain(extended: Boolean): Unit
```

1. Calls `explain` with `extended` flag disabled

`Dataset.explain` is a high-level operator that prints the [logical](#) and (with `extended` flag enabled) [physical](#) plans to the console.

```
val records = spark.
  readStream.
  format("rate").
  load
scala> records.explain
== Physical Plan ==
StreamingRelation rate, [timestamp#0, value#1L]

scala> records.explain(extended = true)
== Parsed Logical Plan ==
StreamingRelation DataSource(org.apache.spark.sql.SparkSession@4071aa13, rate, List(), None, List(), None, Map(), None), rate, [timestamp#0, value#1L]

== Analyzed Logical Plan ==
timestamp: timestamp, value: bigint
StreamingRelation DataSource(org.apache.spark.sql.SparkSession@4071aa13, rate, List(), None, List(), None, Map(), None), rate, [timestamp#0, value#1L]

== Optimized Logical Plan ==
StreamingRelation DataSource(org.apache.spark.sql.SparkSession@4071aa13, rate, List(), None, List(), None, Map(), None), rate, [timestamp#0, value#1L]

== Physical Plan ==
StreamingRelation rate, [timestamp#0, value#1L]
```

Internally, `explain` creates a `ExplainCommand` runnable command with the logical plan and `extended` flag.

`explain` then executes the plan with `ExplainCommand` runnable command and collects the results that are printed out to the standard output.

explain uses `SparkSession` to access the current `SessionState` to execute the plan.

Note

```
import org.apache.spark.sql.execution.command.ExplainCommand
val explain = ExplainCommand(...)
spark.sessionState.executePlan(explain)
```

For streaming Datasets, `ExplainCommand` command simply creates a [IncrementalExecution](#) for the `SparkSession` and the logical plan.

Note

For the purpose of `explain`, `IncrementalExecution` is created with the output mode `Append`, checkpoint location `<unknown>`, run id a random number, current batch id `0` and offset metadata empty. They do not really matter when explaining the load-part of a streaming query.

# groupBy Operator — Untyped Streaming Aggregation (with Implicit State Logic)

```
groupBy(cols: Column*): RelationalGroupedDataset
groupBy(col1: String, cols: String*): RelationalGroupedDataset
```

groupBy operator...FIXME

```
scala> spark.version
res0: String = 2.3.0-SNAPSHOT

// Since I'm with SNAPSHOT
// Remember to remove ~/.ivy2/cache/org.apache.spark
// Make sure that ~/.ivy2/jars/org.apache.spark_spark-sql-kafka-0-10_2.11-2.3.0-SNAPSHOT.jar is the latest
// Start spark-shell as follows
/***
./bin/spark-shell --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.3.0-SNAPSHOT
***/

val fromTopic1 = spark.
  readStream.
  format("kafka").
  option("subscribe", "topic1").
  option("kafka.bootstrap.servers", "localhost:9092").
  load

// extract event time et al
// time,key,value
/*
2017-08-23T00:00:00.002Z,1,now
2017-08-23T00:05:00.002Z,1,5 mins later
2017-08-23T00:09:00.002Z,1,9 mins later
2017-08-23T00:11:00.002Z,1,11 mins later
2017-08-23T01:00:00.002Z,1,1 hour later
// late event = watermark should be (1 hour - 10 minutes) already
2017-08-23T00:49:59.002Z,1,==> SHOULD NOT BE INCLUDED in aggregation as too late <=

CAUTION: FIXME SHOULD NOT BE INCLUDED is included contrary to my understanding?!
*/
val timedValues = fromTopic1.
  select('value cast "string").
  withColumn("tokens", split('value, ",")).
  withColumn("time", to_timestamp('tokens(0))).
  withColumn("key", 'tokens(1) cast "int").
  withColumn("value", 'tokens(2)).
  select("time", "key", "value")
```

```

// aggregation with watermark
val counts = timedValues.
  withWatermark("time", "10 minutes").
  groupBy("key").
  agg(collect_list('value) as "values", collect_list('time) as "times")

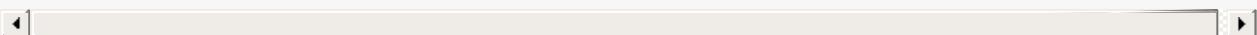
// Note that StatefulOperatorStateInfo is mostly generic
// since no batch-specific values are currently available
// only after the first streaming batch
scala> counts.explain
== Physical Plan ==
ObjectHashAggregate(keys=[key#27], functions=[collect_list(value#33, 0, 0), collect_li
st(time#22-T600000ms, 0, 0)])
+- Exchange hashpartitioning(key#27, 200)
  +- StateStoreSave [key#27], StatefulOperatorStateInfo(<unknown>, 25149816-1f14-4901-
af13-896286a26d42,0,0), Append, 0
    +- ObjectHashAggregate(keys=[key#27], functions=[merge_collect_list(value#33, 0,
0), merge_collect_list(time#22-T600000ms, 0, 0)])
      +- Exchange hashpartitioning(key#27, 200)
        +- StateStoreRestore [key#27], StatefulOperatorStateInfo(<unknown>, 25149816-
1f14-4901-af13-896286a26d42,0,0)
          +- ObjectHashAggregate(keys=[key#27], functions=[merge_collect_list(val
ue#33, 0, 0), merge_collect_list(time#22-T600000ms, 0, 0)])
            +- Exchange hashpartitioning(key#27, 200)
              +- ObjectHashAggregate(keys=[key#27], functions=[partial_collect_
list(value#33, 0, 0), partial_collect_list(time#22-T600000ms, 0, 0)])
                +- EventTimeWatermark time#22: timestamp, interval 10 minutes
                  +- *Project [cast(split(cast(value#1 as string), ,)[0] as t
imestamp) AS time#22, cast(split(cast(value#1 as string), ,)[1] as int) AS key#27, spl
it(cast(value#1 as string), ,)[2] AS value#33]
                    +- StreamingRelation kafka, [key#0, value#1, topic#2, pa
rtition#3, offset#4L, timestamp#5, timestampType#6]

import org.apache.spark.sql.streaming.-
import scala.concurrent.duration.-
val sq = counts.writeStream.
  format("console").
  option("truncate", false).
  trigger(Trigger.ProcessingTime(30.seconds)).
  outputMode(OutputMode.Update). // <-- only Update or Complete acceptable because of
groupBy aggregation
  start

// After StreamingQuery was started,
// the physical plan is complete (with batch-specific values)
scala> sq.explain
== Physical Plan ==
ObjectHashAggregate(keys=[key#27], functions=[collect_list(value#33, 0, 0), collect_li
st(time#22-T600000ms, 0, 0)])
+- Exchange hashpartitioning(key#27, 200)
  +- StateStoreSave [key#27], StatefulOperatorStateInfo(file:/private/var/folders/0w/
kb0d3rqn4zb9fcc91pxhgn8w0000gn/T/temporary-635d6519-b6ca-4686-9b6b-5db0e83cf51/state,

```

```
855cec1c-25dc-4a86-ae54-c6cdd4ed02ec, 0, 0), Update, 0
    +- ObjectHashAggregate(keys=[key#27], functions=[merge_collect_list(value#33, 0,
0), merge_collect_list(time#22-T600000ms, 0, 0)])
        +- Exchange hashpartitioning(key#27, 200)
            +- StateStoreRestore [key#27], StatefulOperatorStateInfo(file:/private/var
/folders/0w/kb0d3rqn4zb9fcc91pxhgn8w0000gn/T/temporary-635d6519-b6ca-4686-9b6b-5db0e83
cf51/state,855cec1c-25dc-4a86-ae54-c6cdd4ed02ec, 0, 0)
                +- ObjectHashAggregate(keys=[key#27], functions=[merge_collect_list(val
ue#33, 0, 0), merge_collect_list(time#22-T600000ms, 0, 0)])
                    +- Exchange hashpartitioning(key#27, 200)
                        +- ObjectHashAggregate(keys=[key#27], functions=[partial_collect_
list(value#33, 0, 0), partial_collect_list(time#22-T600000ms, 0, 0)])
                            +- EventTimeWatermark time#22: timestamp, interval 10 minutes
                                +- *Project [cast(split(cast(value#76 as string), ,)[0] as
timestamp) AS time#22, cast(split(cast(value#76 as string), ,)[1] as int) AS key#27,
split(cast(value#76 as string), ,)[2] AS value#33]
                                    +- Scan ExistingRDD[key#75,value#76,topic#77,partition#78
,offset#79L,timestamp#80,timestampType#81]
```



# groupByKey Operator — Streaming Aggregation

- [Introduction](#)
- [Example: Aggregating Orders Per Zip Code](#)
- [Example: Aggregating Metrics Per Device](#)

## Introduction

```
groupByKey[K: Encoder](func: T => K): KeyValueGroupedDataset[K, T]
```

`groupByKey` operator creates a [KeyValueGroupedDataset](#) (with keys of type `K` and rows of type `T`) to apply aggregation functions over groups of rows (of type `T`) by key (of type `K`) per the given `func` key-generating function.

Note	The type of the input argument of <code>func</code> is the type of rows in the Dataset (i.e. <code>Dataset[T]</code> ).
------	---

`groupByKey` simply applies the `func` function to every row (of type `T`) and associates it with a logical group per key (of type `K`).

```
func: T => K
```

Internally, `groupByKey` creates a structured query with the `AppendColumns` unary logical operator (with the given `func` and the analyzed logical plan of the target `Dataset` that `groupByKey` was executed on) and creates a new `QueryExecution`.

In the end, `groupByKey` creates a [KeyValueGroupedDataset](#) with the following:

- Encoders for `K` keys and `T` rows
- The new `QueryExecution` (with the `AppendColumns` unary logical operator)
- The output schema of the analyzed logical plan
- The new columns of the `AppendColumns` logical operator (i.e. the attributes of the key)

```

scala> :type sq
org.apache.spark.sql.Dataset[Long]

val baseCode = 'A'.toInt
val byUpperChar = (n: java.lang.Long) => (n % 3 + baseCode).toString
val kvs = sq.groupByKey(byUpperChar)

scala> :type kvs
org.apache.spark.sql.KeyValueGroupedDataset[String, Long]

// Peeking under the surface of KeyValueGroupedDataset
import org.apache.spark.sql.catalyst.plans.logical.AppendColumns
val appendColumnsOp = kvs.queryExecution.analyzed.collect { case ac: AppendColumns =>
  ac }.head
scala> println(appendColumnsOp.newColumns)
List(value#7)

```

## Example: Aggregating Orders Per Zip Code

Go to [Demo: groupByKey Streaming Aggregation in Update Mode](#).

## Example: Aggregating Metrics Per Device

The following example code shows how to apply `groupByKey` operator to a structured stream of timestamped values of different devices.

```

// input stream
import java.sql.Timestamp
val signals = spark.
  readStream.
  format("rate").
  option("rowsPerSecond", 1).
  load.
  withColumn("value", $"value" % 10) // <-- randomize the values (just for fun)
  withColumn("deviceId", lit(util.Random.nextInt(10))). // <-- 10 devices randomly assigned to values
  as[(Timestamp, Long, Int)] // <-- convert to a "better" type (from "unpleasant" Row)

// stream processing using groupByKey operator
// groupByKey(func: ((Timestamp, Long, Int)) => K): KeyValueGroupedDataset[K, (Timestamp, Long, Int)]
// K becomes Int which is a device id
val deviceId: ((Timestamp, Long, Int)) => Int = { case (_, _, deviceId) => deviceId }
scala> val signalsByDevice = signals.groupByKey(deviceId)
signalsByDevice: org.apache.spark.sql.KeyValueGroupedDataset[Int,(java.sql.Timestamp, Long, Int)] = org.apache.spark.sql.KeyValueGroupedDataset@19d40bc6

```



# withWatermark Operator — Event-Time Watermark

```
withWatermark(eventTime: String, delayThreshold: String): Dataset[T]
```

`withWatermark` specifies the `eventTime` column for **event time watermark** and `delayThreshold` for **event lateness**.

`eventTime` specifies the column to use for watermark and can be either part of `Dataset` from the source or custom-generated using `current_time` or `current_timestamp` functions.

Note	<p><b>Watermark</b> tracks a point in time before which it is assumed no more late events are supposed to arrive (and if they have, the late events are considered really late and simply dropped).</p>
------	---

Note	<p>Spark Structured Streaming uses watermark for the following:</p> <ul style="list-style-type: none"><li>• To know when a given time window aggregation (using <a href="#">groupBy</a> operator with <a href="#">window</a> function) can be finalized and thus emitted when using output modes that do not allow updates, like <a href="#">Append</a> output mode.</li><li>• To minimize the amount of state that we need to keep for ongoing aggregations, e.g. <a href="#">mapGroupsWithState</a> (for implicit state management), <a href="#">flatMapGroupsWithState</a> (for user-defined state management) and <a href="#">dropDuplicates</a> operators.</li></ul>
------	---

The **current watermark** is computed by looking at the maximum `eventTime` seen across all of the partitions in a query minus a user-specified `delayThreshold`. Due to the cost of coordinating this value across partitions, the actual watermark used is only guaranteed to be at least `delayThreshold` behind the actual event time.

Note	<p>In some cases Spark may still process records that arrive more than <code>delayThreshold</code> late.</p>
------	--

# window Function — Stream Time Windows

`window` is a standard function that generates **tumbling**, **sliding** or **delayed** stream time window ranges (on a timestamp column).

```
window(
    timeColumn: Column,
    windowDuration: String): Column  (1)
window(
    timeColumn: Column,
    windowDuration: String,
    slideDuration: String): Column   (2)
window(
    timeColumn: Column,
    windowDuration: String,
    slideDuration: String,
    startTime: String): Column       (3)
```

1. Creates a tumbling time window with `slideDuration` as `windowDuration` and `0` second for `startTime`
2. Creates a sliding time window with `0` second for `startTime`
3. Creates a delayed time window

	From <a href="#">Tumbling Window (Azure Stream Analytics)</a> :
Note	<ul style="list-style-type: none"> <li>■ <b>Tumbling windows</b> are a series of fixed-sized, non-overlapping and contiguous time intervals.</li> </ul>

	From <a href="#">Introducing Stream Windows in Apache Flink</a> :
Note	<ul style="list-style-type: none"> <li>■ <b>Tumbling windows</b> group elements of a stream into finite sets where each set corresponds to an interval.</li> <li>■ <b>Tumbling windows</b> discretize a stream into non-overlapping windows.</li> </ul>

```
scala> val timeColumn = window($"time", "5 seconds")
timeColumn: org.apache.spark.sql.Column = timewindow(time, 5000000, 5000000, 0) AS `window`
```

`timeColumn` should be of `TimestampType`, i.e. with [java.sql.Timestamp](#) values.

Tip	Use <a href="#">java.sql.Timestamp.from</a> or <a href="#">java.sql.Timestamp.valueOf</a> factory methods to create <code>Timestamp</code> instances.
-----	---

```
// https://docs.oracle.com/javase/8/docs/api/java/time/LocalDateTime.html
import java.time.LocalDateTime
// https://docs.oracle.com/javase/8/docs/api/java/sql/Timestamp.html
import java.sql.Timestamp
val levels = Seq(
    // (year, month, dayOfMonth, hour, minute, second)
    ((2012, 12, 12, 12, 12, 12), 5),
    ((2012, 12, 12, 12, 12, 14), 9),
    ((2012, 12, 12, 13, 13, 14), 4),
    ((2016, 8, 13, 0, 0, 0), 10),
    ((2017, 5, 27, 0, 0, 0), 15)).
    map { case ((yy, mm, dd, h, m, s), a) => (LocalDateTime.of(yy, mm, dd, h, m, s), a)}
).
    map { case (ts, a) => (Timestamp.valueOf(ts), a) }.
    toDF("time", "level")
scala> levels.show
+-----+-----+
|          time|level|
+-----+-----+
|2012-12-12 12:12:12|    5|
|2012-12-12 12:12:14|    9|
|2012-12-12 13:13:14|    4|
|2016-08-13 00:00:00|   10|
|2017-05-27 00:00:00|   15|
+-----+-----+

val q = levels.select(window($"time", "5 seconds"), $"level")
scala> q.show(truncate = false)
+-----+-----+
|window                               |level|
+-----+-----+
|[2012-12-12 12:12:10.0,2012-12-12 12:12:15.0]|5    |
|[2012-12-12 12:12:10.0,2012-12-12 12:12:15.0]|9    |
|[2012-12-12 13:13:10.0,2012-12-12 13:13:15.0]|4    |
|[2016-08-13 00:00:00.0,2016-08-13 00:00:05.0]|10   |
|[2017-05-27 00:00:00.0,2017-05-27 00:00:05.0]|15   |
+-----+-----+

scala> q.printSchema
root
|-- window: struct (nullable = true)
|   |-- start: timestamp (nullable = true)
|   |-- end: timestamp (nullable = true)
|-- level: integer (nullable = false)

// calculating the sum of levels every 5 seconds
val sums = levels.
    groupBy(window($"time", "5 seconds")).
    agg(sum("level") as "level_sum").
    select("window.start", "window.end", "level_sum")
scala> sums.show
+-----+-----+-----+
```

	start	end	level_sum
	2012-12-12 13:13:10	2012-12-12 13:13:15	4
	2012-12-12 12:12:10	2012-12-12 12:12:15	14
	2016-08-13 00:00:00	2016-08-13 00:00:05	10
	2017-05-27 00:00:00	2017-05-27 00:00:05	15

`windowDuration` and `slideDuration` are strings specifying the width of the window for duration and sliding identifiers, respectively.

**Tip**

Use `CalendarInterval` for valid window identifiers.

There are a couple of rules governing the durations:

1. The window duration must be greater than 0
2. The slide duration must be greater than 0.
3. The start time must be greater than or equal to 0.
4. The slide duration must be less than or equal to the window duration.
5. The start time must be less than the slide duration.

**Note**

Only one `window` expression is supported in a query.

**Note**

`null` values are filtered out in `window` expression.

Internally, `window` creates a [Column](#) with `TimeWindow` Catalyst expression under `window` alias.

```
scala> val timeColumn = window($"time", "5 seconds")
timeColumn: org.apache.spark.sql.Column = timewindow(time, 5000000, 5000000, 0) AS `window`  
  

val windowExpr = timeColumn.expr
scala> println(windowExpr.numberedTreeString)
00 timewindow('time, 5000000, 5000000, 0) AS window#23
01 +- timewindow('time, 5000000, 5000000, 0)
02     +- 'time
```

Internally, `TimeWindow` Catalyst expression is simply a struct type with two fields, i.e. `start` and `end`, both of `TimestampType` type.

```

scala> println(windowExpr.dataType)
StructType(StructField(start,.TimestampType,true), StructField(end,.TimestampType,true))

scala> println(windowExpr.dataType.prettyJson)
{
  "type" : "struct",
  "fields" : [ {
    "name" : "start",
    "type" : "timestamp",
    "nullable" : true,
    "metadata" : { }
  }, {
    "name" : "end",
    "type" : "timestamp",
    "nullable" : true,
    "metadata" : { }
  } ]
}

```

**Note**

`TimeWindow` time window Catalyst expression is planned (i.e. *converted*) in `TimeWindowing` logical optimization rule (i.e. `Rule[LogicalPlan]` ) of the Spark SQL logical query plan analyzer.

Find more about the Spark SQL logical query plan analyzer in [Mastering Apache Spark 2](#) gitbook.

## Example — Traffic Sensor

**Note**

The example is borrowed from [Introducing Stream Windows in Apache Flink](#).

The example shows how to use `window` function to model a traffic sensor that counts every 15 seconds the number of vehicles passing a certain location.

# KeyValueGroupedDataset — Streaming Aggregation

`KeyValueGroupedDataset` represents a **grouped dataset** as a result of `Dataset.groupByKey` operator (that aggregates records by a grouping function).

```
// Dataset[T]
groupByKey(func: T => K): KeyValueGroupedDataset[K, T]
```

```
import java.sql.Timestamp
val numGroups = spark.
  readStream.
  format("rate").
  load.
  as[(Timestamp, Long)].
  groupByKey { case (time, value) => value % 2 }

scala> :type numGroups
org.apache.spark.sql.KeyValueGroupedDataset[Long, (java.sql.Timestamp, Long)]
```

`KeyValueGroupedDataset` is also [created](#) for `KeyValueGroupedDataset.keyAs` and `KeyValueGroupedDataset.mapValues` operators.

```
scala> :type numGroups
org.apache.spark.sql.KeyValueGroupedDataset[Long, (java.sql.Timestamp, Long)]

scala> :type numGroups.keyAs[String]
org.apache.spark.sql.KeyValueGroupedDataset[String, (java.sql.Timestamp, Long)]
```

```
scala> :type numGroups
org.apache.spark.sql.KeyValueGroupedDataset[Long, (java.sql.Timestamp, Long)]

val mapped = numGroups.mapValues { case (ts, n) => s"($ts, $n)" }
scala> :type mapped
org.apache.spark.sql.KeyValueGroupedDataset[Long, String]
```

`KeyValueGroupedDataset` works for batch and streaming aggregations, but shines the most when used for [Streaming Aggregation](#).

```

scala> :type numGroups
org.apache.spark.sql.KeyValueGroupedDataset[Long, (java.sql.Timestamp, Long)]


import org.apache.spark.sql.streaming.Trigger
import scala.concurrent.duration._

numGroups.
  mapGroups { case(group, values) => values.size }.
  writeStream.
  format("console").
  trigger(Trigger.ProcessingTime(10.seconds)).
  start

-----
Batch: 0
-----
+---+
| value|
+---+
+---+


-----
Batch: 1
-----
+---+
| value|
+---+
|   3|
|   2|
+---+


-----
Batch: 2
-----
+---+
| value|
+---+
|   5|
|   5|
+---+


// Eventually...
spark.streams.active.foreach(_.stop)

```

The most prestigious use case of `KeyValueGroupedDataset` however is [Arbitrary Stateful Streaming Aggregation](#) that allows for accumulating **streaming state** (by means of `GroupState`) using `mapGroupsWithState` and the more advanced `flatMapGroupsWithState` operators.

Table 1. KeyValueGroupedDataset's Operators

Operator	Description

agg	<pre> agg[U1](col1: TypedColumn[V, U1]): Dataset[(K, U1)] agg[U1, U2](   col1: TypedColumn[V, U1],   col2: TypedColumn[V, U2]): Dataset[(K, U1, U2)] agg[U1, U2, U3](   col1: TypedColumn[V, U1],   col2: TypedColumn[V, U2],   col3: TypedColumn[V, U3]): Dataset[(K, U1, U2, U3)] agg[U1, U2, U3, U4](   col1: TypedColumn[V, U1],   col2: TypedColumn[V, U2],   col3: TypedColumn[V, U3],   col4: TypedColumn[V, U4]): Dataset[(K, U1, U2, U3, U4)] </pre>		
cogroup	<pre>cogroup[U, R : Encoder](   other: KeyValueGroupedDataset[K, U])(f: (K, Iterator[V], Iterator[U]) =&gt; TraversableOnce[R]): Dataset[(K, R)]</pre>		
count	<pre>count(): Dataset[(K, Long)]</pre>		
flatMapGroups	<pre>flatMapGroups[U : Encoder](f: (K, Iterator[V]) =&gt; TraversableOnce[U]): Dataset[(K, U)]</pre>		
flatMapGroupsWithState	<pre>flatMapGroupsWithState[S: Encoder, U: Encoder](   outputMode: OutputMode,   timeoutConf: GroupStateTimeout)(   func: (K, Iterator[V], GroupState[S]) =&gt; Iterator[U]): Dataset[(K, U)]</pre> <p><a href="#">Arbitrary Stateful Streaming Aggregation</a> - streaming aggregation and state timeout</p> <table border="1"> <tr> <td>Note</td><td>The difference between this <code>flatMapGroupsWithState</code> <code>mapGroupsWithState</code> operators is the state function or more elements (that are in turn the rows in the resulting Dataset ).</td></tr> </table>	Note	The difference between this <code>flatMapGroupsWithState</code> <code>mapGroupsWithState</code> operators is the state function or more elements (that are in turn the rows in the resulting Dataset ).
Note	The difference between this <code>flatMapGroupsWithState</code> <code>mapGroupsWithState</code> operators is the state function or more elements (that are in turn the rows in the resulting Dataset ).		
keyAs	<pre>keys: Dataset[K] keyAs[L : Encoder]: KeyValueGroupedDataset[L, V]</pre>		
mapGroups	<pre>mapGroups[U : Encoder](f: (K, Iterator[V]) =&gt; U): Dataset[U]</pre>		
	<pre>mapGroupsWithState[S: Encoder, U: Encoder](   func: (K, Iterator[V], GroupState[S]) =&gt; U): Dataset[U] mapGroupsWithState[S: Encoder, U: Encoder](   timeoutConf: GroupStateTimeout)(   func: (K, Iterator[V], GroupState[S]) =&gt; U): Dataset[U]</pre>		

<code>mapGroupsWithState</code>	Creates a new Dataset with <a href="#">FlatMapGroupsWithState</a> logical  <table border="1"> <tr> <td>Note</td><td>The difference between <code>mapGroupsWithState</code> and <code>flatMapGroupsWithState</code> is the state function that generates exactly one element row in the result dataset).</td></tr> </table>	Note	The difference between <code>mapGroupsWithState</code> and <code>flatMapGroupsWithState</code> is the state function that generates exactly one element row in the result dataset).
Note	The difference between <code>mapGroupsWithState</code> and <code>flatMapGroupsWithState</code> is the state function that generates exactly one element row in the result dataset).		
<code>mapValues</code>	<code>mapValues[W : Encoder](func: V =&gt; W): KeyValueGroupedDataset[W]</code>		
<code>reduceGroups</code>	<code>reduceGroups(f: (V, V) =&gt; V): Dataset[(K, V)]</code>		

## Creating KeyValueGroupedDataset Instance

`KeyValueGroupedDataset` takes the following when created:

- `Encoder` for keys
- `Encoder` for values
- `QueryExecution`
- Data attributes
- Grouping attributes

# mapGroupsWithState Operator — Stateful Streaming Aggregation (with Explicit State Logic)

```
mapGroupsWithState[S: Encoder, U: Encoder](
  func: (K, Iterator[V], GroupState[S]) => U): Dataset[U] (1)
mapGroupsWithState[S: Encoder, U: Encoder](
  timeoutConf: GroupStateTimeout)(
  func: (K, Iterator[V], GroupState[S]) => U): Dataset[U]
```

1. Uses `GroupStateTimeout.NoTimeout` for `timeoutConf`

`mapGroupsWithState` operator...FIXME

Note

`mapGroupsWithState` is a special case of [flatMapGroupsWithState](#) operator with the following:

- `func` being transformed to return a single-element `Iterator`
- [Update](#) output mode

`mapGroupsWithState` also creates a `FlatMapGroupsWithState` with [isMapGroupsWithState](#) internal flag enabled.

```
// numGroups defined at the beginning
scala> :type numGroups
org.apache.spark.sql.KeyValueGroupedDataset[Long, (java.sql.Timestamp, Long)]

import org.apache.spark.sql.streaming.GroupState
def mappingFunc(key: Long, values: Iterator[(java.sql.Timestamp, Long)], state: GroupState[Long]): Long = {
  println(s">>> key: $key => state: $state")
  val newState = state.getOption.map(_ + values.size).getOrElse(0L)
  state.update(newState)
  key
}

import org.apache.spark.sql.streaming.GroupStateTimeout
val longs = numGroups.mapGroupsWithState(
  timeoutConf = GroupStateTimeout.ProcessingTimeTimeout(
    func = mappingFunc)

import org.apache.spark.sql.streaming.{OutputMode, Trigger}
import scala.concurrent.duration._
val q = longs.
  writeStream.
```

## mapGroupsWithState Operator

---

```
format("console").  
trigger(Trigger.ProcessingTime(10.seconds)).  
outputMode(OutputMode.Update). // <-- required for mapGroupsWithState  
start  
  
// Note GroupState  
  
-----  
Batch: 1  
-----  
>>> key: 0 => state: GroupState(<undefined>)  
>>> key: 1 => state: GroupState(<undefined>)  
+---+  
|value|  
+---+  
| 0 |  
| 1 |  
+---+  
  
-----  
Batch: 2  
-----  
>>> key: 0 => state: GroupState(0)  
>>> key: 1 => state: GroupState(0)  
+---+  
|value|  
+---+  
| 0 |  
| 1 |  
+---+  
  
-----  
Batch: 3  
-----  
>>> key: 0 => state: GroupState(4)  
>>> key: 1 => state: GroupState(4)  
+---+  
|value|  
+---+  
| 0 |  
| 1 |  
+---+  
  
// in the end  
spark.streams.active.foreach(_.stop)
```

# flatMapGroupsWithState Operator — Arbitrary Stateful Streaming Aggregation (with Explicit State Logic)

```
KeyValueGroupedDataset[K, V].flatMapGroupsWithState[S: Encoder, U: Encoder](
  outputMode: OutputMode,
  timeoutConf: GroupStateTimeout)(
  func: (K, Iterator[V], GroupState[S]) => Iterator[U]): Dataset[U]
```

`flatMapGroupsWithState` operator is used for [Arbitrary Stateful Streaming Aggregation \(with Explicit State Logic\)](#).

`flatMapGroupsWithState` requires that the given `OutputMode` is either `Append` or `Update` (and reports an `IllegalArgumentException` at runtime).

Note	An <code>OutputMode</code> is a required argument, but does not seem to be used at all. Check out the question <a href="#">What's the purpose of OutputMode in flatMapGroupsWithState? How/where is it used?</a> on StackOverflow.
------	---

Every time the state function `func` is executed for a key, the state (as `GroupState[S]`) is for this key only.

Note	<ul style="list-style-type: none"> <li>• <code>k</code> is the type of the keys in <code>KeyValueGroupedDataset</code></li> <li>• <code>v</code> is the type of the values (per key) in <code>KeyValueGroupedDataset</code></li> <li>• <code>s</code> is the user-defined type of the state as maintained for each group</li> <li>• <code>u</code> is the type of rows in the result <code>Dataset</code></li> </ul>
------	--

Internally, `flatMapGroupsWithState` creates a new `Dataset` with [FlatMapGroupsWithState](#) unary logical operator.

# StreamingQueryManager — Streaming Query Management

`StreamingQueryManager` is the management interface for active streaming queries of a `SparkSession`.

Table 1. `StreamingQueryManager` API

Method	Description
<code>active</code>	<pre>active: Array[StreamingQuery]</pre> <p>Active structured queries</p>
<code>addListener</code>	<pre>addListener(listener: StreamingQueryListener): Unit</pre> <p>Registers (adds) a <code>StreamingQueryListener</code></p>
<code>awaitAnyTermination</code>	<pre>awaitAnyTermination(): Unit awaitAnyTermination(timeoutMs: Long): Boolean</pre> <p>Waits until any streaming query terminates or <code>timeoutMs</code> elapses</p>
<code>get</code>	<pre>get(id: String): StreamingQuery get(id: UUID): StreamingQuery</pre> <p>Gets the <code>StreamingQuery</code> by id</p>
<code>removeListener</code>	<pre>removeListener(   listener: StreamingQueryListener): Unit</pre> <p>De-registers (removes) the <code>StreamingQueryListener</code></p>
<code>resetTerminated</code>	<pre>resetTerminated(): Unit</pre> <p>Resets the internal registry of the terminated streaming queries (that lets <code>awaitAnyTermination</code> to be used again)</p>

`StreamingQueryManager` is available using `SparkSession.streams` property.

```

scala> :type spark
org.apache.spark.sql.SparkSession

scala> :type spark.streams
org.apache.spark.sql.streaming.StreamingQueryManager

```

`StreamingQueryManager` is created when `SessionState` is created.

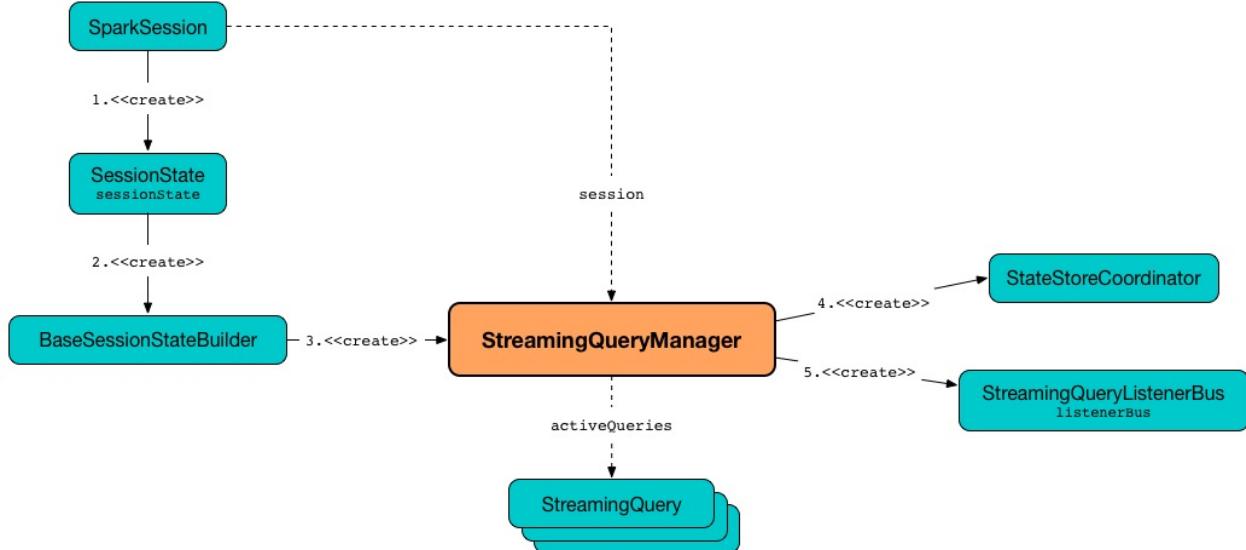


Figure 1. `StreamingQueryManager`

Tip	Read up on <a href="#">SessionState</a> in <a href="#">The Internals of Spark SQL</a> gitbook.
-----	--

`StreamingQueryManager` is used (internally) to create a `StreamingQuery` (and its `StreamExecution`).

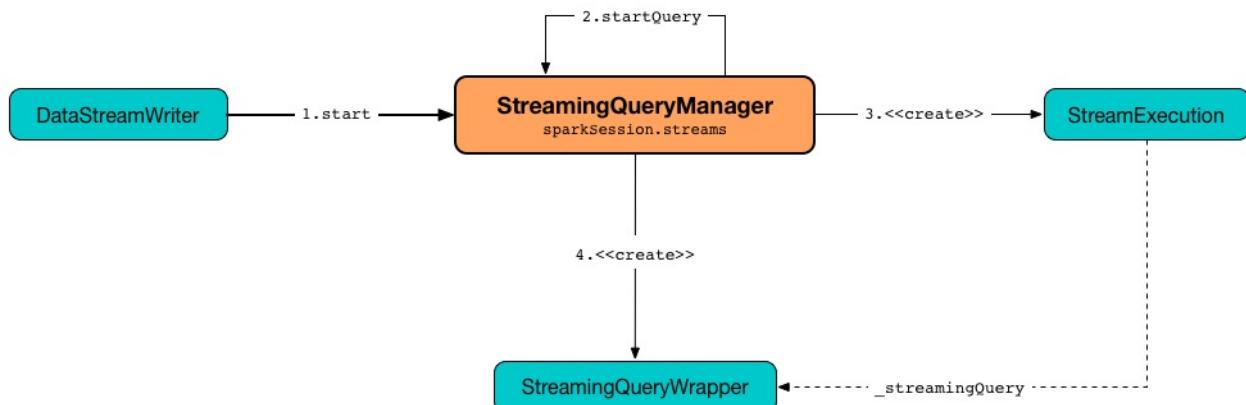


Figure 2. `StreamingQueryManager` Creates `StreamingQuery` (and `StreamExecution`)  
`StreamingQueryManager` is notified about state changes of a structured query and passes them along (to registered listeners).

`StreamingQueryManager` takes a single `SparkSession` when created.

## StreamingQueryListenerBus — `listenerBus` Internal Property

```
listenerBus: StreamingQueryListenerBus
```

`listenerBus` is a [StreamingQueryListenerBus](#) (for the current [SparkSession](#)) that is created immediately when `streamingQueryManager` is [created](#).

`listenerBus` is used for the following:

- [Register](#) or [de-register](#) a given [StreamingQueryListener](#)
- [Post a streaming event](#) (and notify [registered StreamingQueryListeners](#) about the [event](#))

## Getting All Active Streaming Queries — `active` Method

```
active: Array[StreamingQuery]
```

`active` gets [all active streaming queries](#).

## Getting Active Continuous Query By Name — `get` Method

```
get(name: String): StreamingQuery
```

`get` method returns a [StreamingQuery](#) by `name`.

It may throw an `IllegalArgumentException` when no [StreamingQuery](#) exists for the `name`.

```
java.lang.IllegalArgumentException: There is no active query with name hello
  at org.apache.spark.sql.StreamingQueryManager$$anonfun$get$1.apply(StreamingQueryManager.scala:59)
  at org.apache.spark.sql.StreamingQueryManager$$anonfun$get$1.apply(StreamingQueryManager.scala:59)
  at scala.collection.MapLike$class.getOrElse(MapLike.scala:128)
  at scala.collection.AbstractMap.getOrElse(Map.scala:59)
  at org.apache.spark.sql.StreamingQueryManager.get(StreamingQueryManager.scala:58)
  ... 49 elided
```

## Registering StreamingQueryListener — `addListener` Method

```
addListener(listener: StreamingQueryListener): Unit
```

`addListener` requests the `StreamingQueryListenerBus` to `add` the input `listener`.

## De-Registering StreamingQueryListener — `removeListener` Method

```
removeListener(listener: StreamingQueryListener): Unit
```

`removeListener` requests `StreamingQueryListenerBus` to `remove` the input `listener`.

## Waiting for Any Streaming Query Termination — `awaitAnyTermination` Method

```
awaitAnyTermination(): Unit
awaitAnyTermination(timeoutMs: Long): Boolean
```

`awaitAnyTermination` acquires a lock on `awaitTerminationLock` and waits until any streaming query has finished (i.e. `lastTerminatedQuery` is available) or `timeoutMs` has expired.

`awaitAnyTermination` re-throws the `StreamingQueryException` from `lastTerminatedQuery` if it reported one.

## `resetTerminated` Method

```
resetTerminated(): Unit
```

`resetTerminated` forgets about the past-terminated query (so that `awaitAnyTermination` can be used again to wait for a new streaming query termination).

Internally, `resetTerminated` acquires a lock on `awaitTerminationLock` and simply resets `lastTerminatedQuery` (i.e. sets it to `null`).

## Creating Streaming Query — `createQuery` Internal Method

```
createQuery(  
    userSpecifiedName: Option[String],  
    userSpecifiedCheckpointLocation: Option[String],  
    df: DataFrame,  
    extraOptions: Map[String, String],  
    sink: BaseStreamingSink,  
    outputMode: OutputMode,  
    useTempCheckpointLocation: Boolean,  
    recoverFromCheckpointLocation: Boolean,  
    trigger: Trigger,  
    triggerClock: Clock): StreamingQueryWrapper
```

`createQuery` creates a [StreamingQueryWrapper](#) (for a [StreamExecution](#) per the input user-defined properties).

Internally, `createQuery` first finds the name of the checkpoint directory of a query (aka **checkpoint location**) in the following order:

1. Exactly the input `userSpecifiedCheckpointLocation` if defined
2. [spark.sql.streaming.checkpointLocation](#) Spark property if defined for the parent directory with a subdirectory per the optional `userSpecifiedName` (or a randomly-generated UUID)
3. (only when `useTempCheckpointLocation` is enabled) A temporary directory (as specified by `java.io.tmpdir` JVM property) with a subdirectory with `temporary` prefix.

**Note**

`userSpecifiedCheckpointLocation` can be any path that is acceptable by [Hadoop's Path](#).

If the directory name for the checkpoint location could not be found, `createQuery` reports a [AnalysisException](#).

```
checkpointLocation must be specified either through option("checkpointLocation", ...)  
or SparkSession.conf.set("spark.sql.streaming.checkpointLocation", ...)
```

`createQuery` reports a [AnalysisException](#) when the input `recoverFromCheckpointLocation` flag is turned off but there is **offsets** directory in the checkpoint location.

`createQuery` makes sure that the logical plan of the structured query is analyzed (i.e. no logical errors have been found).

Unless [spark.sql.streaming.unsupportedOperationCheck](#) Spark property is turned on, `createQuery` checks the logical plan of the streaming query for unsupported operations.

(only when `spark.sql.adaptive.enabled` Spark property is turned on) `createQuery` prints out a WARN message to the logs:

```
WARN spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
```

In the end, `createQuery` creates a [StreamingQueryWrapper](#) with a new [MicroBatchExecution](#).

#### Note

`recoverFromCheckpointLocation` flag corresponds to `recoverFromCheckpointLocation` flag that `streamingQueryManager` uses to [start a streaming query](#) and which is enabled by default (and is in fact the only place where `createQuery` is used).

- `memory` sink has the flag enabled for [Complete](#) output mode only
- `foreach` sink has the flag always enabled
- `console` sink has the flag always disabled
- all other sinks have the flag always enabled

#### Note

`userSpecifiedName` corresponds to `queryName` option (that can be defined using `DataStreamWriter`'s [queryName](#) method) while `userSpecifiedCheckpointLocation` is `checkpointLocation` option.

#### Note

`createQuery` is used exclusively when `streamingQueryManager` is requested to [start a streaming query](#) (when `DataStreamWriter` is requested to [start an execution of a streaming query](#)).

## Starting Streaming Query Execution— `startQuery` Internal Method

```
startQuery(
    userSpecifiedName: Option[String],
    userSpecifiedCheckpointLocation: Option[String],
    df: DataFrame,
    extraOptions: Map[String, String],
    sink: BaseStreamingSink,
    outputMode: OutputMode,
    useTempCheckpointLocation: Boolean = false,
    recoverFromCheckpointLocation: Boolean = true,
    trigger: Trigger = ProcessingTime(0),
    triggerClock: Clock = new SystemClock()): StreamingQuery
```

`startQuery` starts a [streaming query](#) and returns a handle to it.

Note	<code>trigger</code> defaults to <code>0</code> milliseconds (as <code>ProcessingTime(0)</code> ).
------	--

Internally, `startQuery` first creates a `StreamingQueryWrapper`, registers it in `activeQueries` internal registry (by the `id`), requests it for the underlying `StreamExecution` and starts it.

In the end, `startQuery` returns the `StreamingQueryWrapper` (as part of the fluent API so you can chain operators) or throws the exception that was reported when attempting to start the query.

`startQuery` throws an `IllegalArgumentException` when there is another query registered under `name`. `startQuery` looks it up in the `activeQueries` internal registry.

Cannot start query with name [name] as a query with that name is already active
---

`startQuery` throws an `IllegalStateException` when a query is started again from checkpoint. `startQuery` looks it up in `activeQueries` internal registry.

Cannot start query with id [id] as another query with same id is already active. Perhaps you are attempting to restart a query from checkpoint that is already active.
--

Note	<code>startQuery</code> is used exclusively when <code>DataStreamWriter</code> is requested to <a href="#">start an execution of the streaming query</a> .
------	--

## Posting StreamingQueryListener Event to StreamingQueryListenerBus — `postListenerEvent` Internal Method

<code>postListenerEvent(event: StreamingQueryListener.Event): Unit</code>
---

`postListenerEvent` simply posts the input `event` to the internal `event bus` for streaming events (`StreamingQueryListenerBus`).

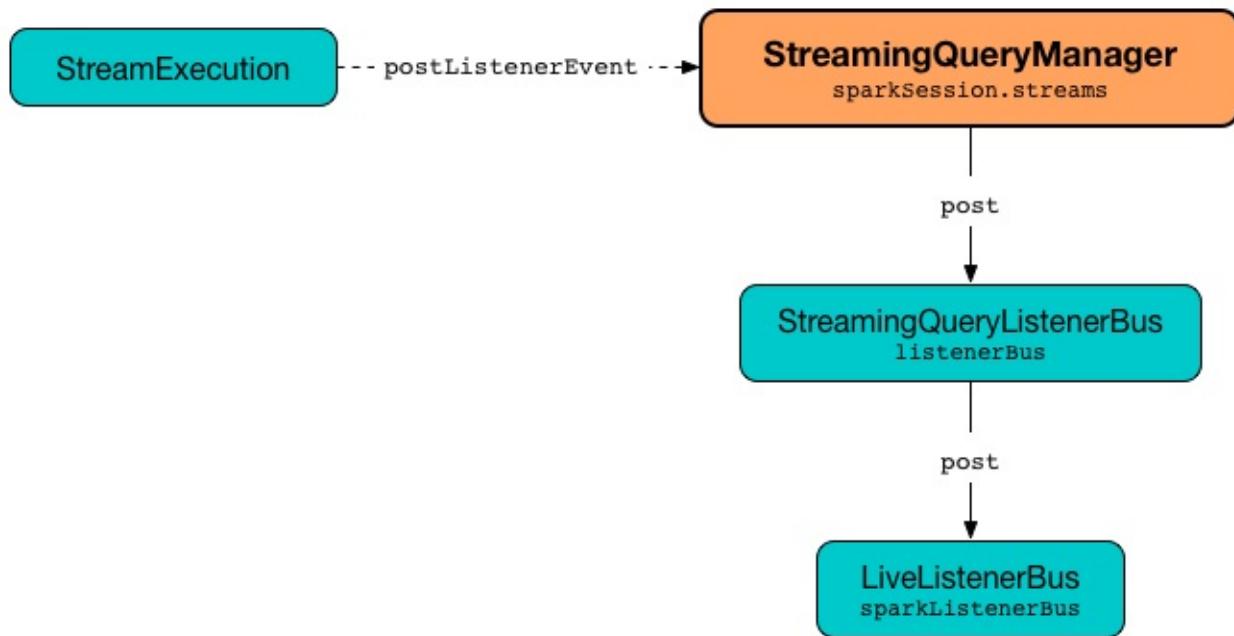


Figure 3. StreamingQueryManager Propagates StreamingQueryListener Events

Note	<code>postListenerEvent</code> is used exclusively when <code>StreamExecution</code> is requested to post a streaming event.
------	--

## Handling Termination of Streaming Query (and Deactivating Query in StateStoreCoordinator)

### — `notifyQueryTermination` Internal Method

```
notifyQueryTermination(terminatedQuery: StreamingQuery): Unit
```

`notifyQueryTermination` removes the `terminatedQuery` from `activeQueries` internal registry (by the `query id`).

`notifyQueryTermination` records the `terminatedQuery` in `lastTerminatedQuery` internal registry (when no earlier streaming query was recorded or the `terminatedQuery` terminated due to an exception).

`notifyQueryTermination` notifies others that are blocked on `awaitTerminationLock`.

In the end, `notifyQueryTermination` requests `StateStoreCoordinator` to deactivate all active runs of the streaming query.

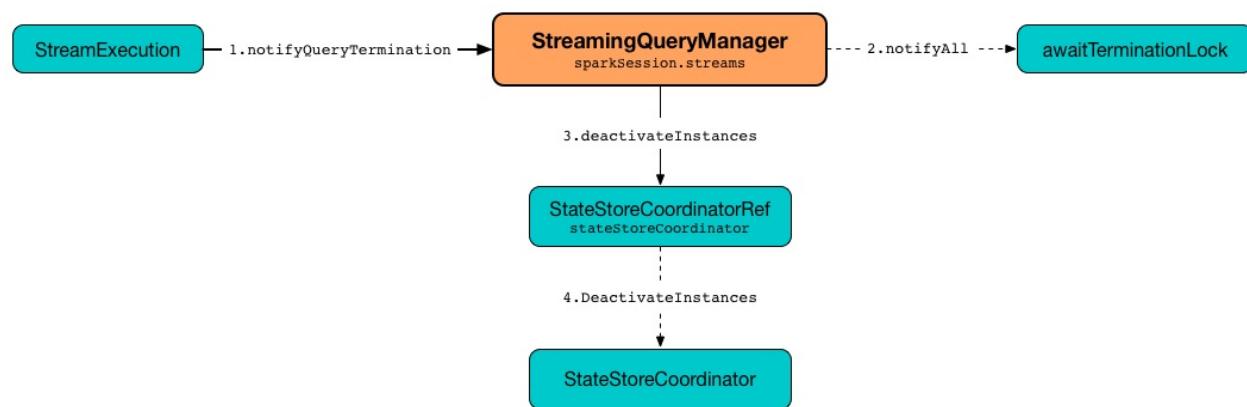


Figure 4. StreamingQueryManager's Marking Streaming Query as Terminated

**Note**

`notifyQueryTermination` is used exclusively when `StreamExecution` is requested to [run a streaming query](#) and the query [has finished \(running streaming batches\)](#) (with or without an exception).

## Internal Properties

Name	Description
activeQueries	<p>Registry of <a href="#">StreamingQueries</a> per <code>UUID</code></p> <p>Used when <code>StreamingQueryManager</code> is requested for <a href="#">active streaming queries</a>, <a href="#">get a streaming query by id</a>, starts a <a href="#">streaming query</a> and <a href="#">is notified that a streaming query has terminated</a>.</p>
activeQueriesLock	
awaitTerminationLock	
lastTerminatedQuery	<p><a href="#">StreamingQuery</a> that has recently been terminated, i.e. <a href="#">stopped</a> or <a href="#">due to an exception</a>.</p> <p><code>null</code> when no streaming query has terminated yet or <a href="#">resetTerminated</a>.</p> <ul style="list-style-type: none"> <li>Used in <a href="#">awaitAnyTermination</a> to know when a streaming query has terminated</li> <li>Set when <code>StreamingQueryManager</code> <a href="#">is notified that a streaming query has terminated</a></li> </ul>
stateStoreCoordinator	<p><a href="#">StateStoreCoordinatorRef</a> to the <code>StateStoreCoordinator</code> RPC Endpoint</p> <ul style="list-style-type: none"> <li><a href="#">Created</a> when <code>StreamingQueryManager</code> <a href="#">is created</a></li> </ul> <p>Used when:</p> <ul style="list-style-type: none"> <li><code>StreamingQueryManager</code> <a href="#">is notified that a streaming query has terminated</a></li> <li>Stateful operators are executed, i.e. <a href="#">FlatMapGroupsWithStateExec</a>, <a href="#">StateStoreRestoreExec</a>, <a href="#">StateStoreSaveExec</a>, <a href="#">StreamingDeduplicateExec</a> and <a href="#">StreamingSymmetricHashJoinExec</a></li> <li>Creating <code>StateStoreRDD</code> (with <code>storeUpdateFunction</code> aborting <code>StateStore</code> when a task fails)</li> </ul>

# SQLConf—Internal Configuration Store

`SQLConf` is an **internal key-value configuration store** for parameters and hints used to configure a Spark Structured Streaming application (and Spark SQL applications in general).

The parameters and hints are accessible as [property accessor methods](#).

`SQLConf` is available as the `conf` property of the `SessionState` of a `SparkSession`.

```
scala> :type spark
org.apache.spark.sql.SparkSession

scala> :type spark.sessionState.conf
org.apache.spark.sql.internal.SQLConf
```

Table 1. SQLConf's Property Accessor Methods

Method Name / Property	Description
<code>continuousStreamingExecutorQueueSize</code> <a href="#">spark.sql.streaming.continuous.executorQueueSize</a>	Used when: <ul style="list-style-type: none"> <li><code>DataSourceV2Scan</code> physical operator the input RDDs (<a href="#">ContinuousDataSource</a>)</li> <li><code>ContinuousCoales</code> physical operator <code>execute</code></li> </ul>
<code>continuousStreamingExecutorPollIntervalMs</code> <a href="#">spark.sql.streaming.continuous.executorPollIntervalMs</a>	Used exclusively when <code>DataSourceV2ScanExec</code> operator is requested for RDDs (and creates a <a href="#">ContinuousDataSource</a> )
<code>disabledV2StreamingMicroBatchReaders</code> <a href="#">spark.sql.streaming.disabledV2MicroBatchReaders</a>	Used exclusively when <code>MicroBatchExecution</code> the <a href="#">analyzed logical plan</a> of a streaming query)
<code>fileSourceLogDeletion</code> <a href="#">spark.sql.streaming.fileSource.log.deletion</a>	Used exclusively when <code>FileStreamSourceLog</code> the <code>isDeletingExpiredFile</code>
<code>fileSourceLogCleanupDelay</code> <a href="#">spark.sql.streaming.fileSource.log.cleanupDelay</a>	Used exclusively when <code>FileStreamSourceLog</code> the <code>fileCleanupDelay</code>

<pre>fileSourceLogCompactInterval</pre> <p><a href="#">spark.sql.streaming.fileSource.log.compactInterval</a></p>	<p>Used exclusively when <a href="#">FileStreamSourceLog</a> is used and the <a href="#">default compaction</a> is used.</p>
<pre>FLATMAPGROUPSWITHSTATE_STATE_FORMAT_VERSION</pre> <p><a href="#">spark.sql.streaming.flatMapGroupsWithState.stateFormatVersion</a></p>	<p>Used when:</p> <ul style="list-style-type: none"> <li>• <a href="#">FlatMapGroupsWithState</a> execution plan is requested to plan a query (and creates <a href="#">FlatMapGroupsWithState</a> physical operator <a href="#">FlatMapGroupsWithState</a> operator)</li> <li>• Among the <a href="#">check properties</a></li> </ul>
<pre>minBatchesToRetain</pre> <p><a href="#">spark.sql.streaming.minBatchesToRetain</a></p>	<p>Used when:</p> <ul style="list-style-type: none"> <li>• <a href="#">CompactibleFileSets</a> are created</li> <li>• <a href="#">StreamExecution</a></li> <li>• <a href="#">StateStoreConf</a> is used</li> </ul>
<pre>SHUFFLE_PARTITIONS</pre> <p><a href="#">spark.sql.shuffle.partitions</a></p>	<p>See <a href="#">spark.sql.shuffle.partitions</a>. Internals of Spark SQL</p>
<pre>stateStoreMinDeltasForSnapshot</pre> <p><a href="#">spark.sql.streaming.stateStore.minDeltasForSnapshot</a></p>	<p>Used (as <a href="#">StateStoreConf.minDeltasForSnapshot</a>) exclusively when <a href="#">HDFSBackedStateStore</a> is used and requested to <a href="#">doSnapshot</a>.</p>
<pre>stateStoreProviderClass</pre> <p><a href="#">spark.sql.streaming.stateStore.providerClass</a></p>	<p>Used when:</p> <ul style="list-style-type: none"> <li>• <a href="#">StateStoreWriter</a>, <a href="#">StateStoreCustomWriter</a>, <a href="#">StateStoreWriter</a> and <a href="#">the metrics and guarantees</a></li> <li>• <a href="#">StateStoreConf</a> is used</li> </ul>
<pre>STREAMING_AGGREGATION_STATE_FORMAT_VERSION</pre> <p><a href="#">spark.sql.streaming.aggregation.stateFormatVersion</a></p>	<p>Used when:</p> <ul style="list-style-type: none"> <li>• <a href="#">StatefulAggregation</a> execution plan is requested to be executed</li> </ul>

	<ul style="list-style-type: none"> <li>• <code>offsetSeqMetadata</code> for the <code>relevantSQLConf</code></li> </ul>
<code>STREAMING_CHECKPOINT_FILE_MANAGER_CLASS</code> <code>spark.sql.streaming.checkpointFileManagerClass</code>	Used exclusively when <code>CheckpointFileManager</code> is requested to <code>create</code> a <code>CheckpointFileManager</code>
<code>streamingMetricsEnabled</code> <code>spark.sql.streaming.metricsEnabled</code>	Used exclusively when <code>StreamExecution</code> is requested to <code>runStream</code> (to control whether to register a <code>metrics repository</code> for a streaming query)
<code>STREAMING_MULTIPLE_WATERMARK_POLICY</code> <code>spark.sql.streaming.multipleWatermarkPolicy</code>	
<code>streamingNoDataMicroBatchesEnabled</code> <code>spark.sql.streaming.noDataMicroBatches.enabled</code>	Used exclusively when <code>MicroBatchExecution</code> engine is requested to run a streaming query
<code>streamingNoDataProgressEventInterval</code> <code>spark.sql.streaming.noDataProgressEventInterval</code>	Used exclusively for Flow API
<code>streamingPollingDelay</code> <code>spark.sql.streaming.pollingDelay</code>	Used exclusively when <code>StreamExecution</code> is created
<code>streamingProgressRetention</code> <code>spark.sql.streaming.numRecentProgressUpdates</code>	Used exclusively when <code>ProgressReporter</code> is requested to update progress of streams (and possibly remove them)

# Configuration Properties

Configuration properties are used to fine-tune Spark Structured Streaming applications.

You can set them for a `SparkSession` when it is created using `config` method.

```
import org.apache.spark.sql.SparkSession
val spark = SparkSession
  .builder
  .config("spark.sql.streaming.metricsEnabled", true)
  .getOrCreate
```

Tip

Read up on [SparkSession](#) in [The Internals of Spark SQL](#) book.

Table 1. Structured Streaming's Properties

Name	Description
<code>spark.sql.streaming.aggregation.stateFormatVersion</code>	<p>(internal) Version of the state format.</p> <p>Default: 2</p> <p>Supported values:</p> <ul style="list-style-type: none"> <li>1 (for the legacy <a href="#">StreamingAggregation</a>)</li> <li>2 (for the default <a href="#">StreamingAggregation</a>)</li> </ul> <p>Used when <a href="#">StatefulAggregation</a> planning strategy is executed on a streaming query with a <code>stateFormat</code> that boils down to creating a <code>StreamingAggregation</code> with the proper <i>implementation</i> of <a href="#">StreamingAggregation</a>.</p> <p>Among the <a href="#">checkpointing</a> strategies, this one is supposed to be overridden by the <code>stateFormat</code> if has once been started. It is used to restore from a checkpoint after a failure.</p>
<code>spark.sql.streaming.checkpointFileManagerClass</code>	<p>(internal) <a href="#">CheckpointFileManager</a> class used to manage checkpoint files atomicity.</p> <p>Default: <a href="#">FileContextBasedCheckpointFileManager</a> (with <a href="#">FileSystemBased</a> implementation in case of unsupported file systems or metadata files)</p>

<code>spark.sql.streaming.checkpointLocation</code>	Default checkpoint directory Default: (empty)
<code>spark.sql.streaming.continuous.executorQueueSize</code>	(internal) The size (max number of partitions) of the queue used in continuous execution to store the results of a Continuous Query. Default: 1024
<code>spark.sql.streaming.continuous.executorPollIntervalMs</code>	(internal) The interval (in ms) between consecutive continuous execution runs. It is triggered when whether the epoch has finished or not. Default: 100 (ms)
<code>spark.sql.streaming.disabledV2MicroBatchReaders</code>	(internal) A comma-separated list of class names of data sources that do not support <a href="#">MicroBatchReadSupport</a> . If these sources will fall back to <a href="#">BatchReadSupport</a> . Default: (empty)  Use <a href="#">SQLConf.disabledV2StreamingReaders</a> to get the current value.
<code>spark.sql.streaming.fileSource.log.cleanupDelay</code>	(internal) How long (in minutes) a file needs to be visible for all readers to see it. Default: 10 (minutes)  Use <a href="#">SQLConf.fileSourceLogCleanupDelay</a> to get the current value.
<code>spark.sql.streaming.fileSource.log.compactInterval</code>	(internal) Number of log files to keep before compacting. Previous files are compressed. Default: 10  Must be a positive value.  Use <a href="#">SQLConf.fileSourceLogCompactInterval</a> to get the current value.
<code>spark.sql.streaming.fileSource.log.deletion</code>	(internal) Whether to delete old log files for a file stream source. Default: true  Use <a href="#">SQLConf.fileSourceLogDeletion</a> to get the current value.

	<p><b>(internal)</b> State format StateManager for FlatN physical operator</p> <p>Default: 2</p> <p>Supported values:</p> <ul style="list-style-type: none"> <li>• 1</li> <li>• 2</li> </ul> <p>Among the checkpoints supposed to be overridden has once been started from a checkpoint after</p>
	<p><b>(internal)</b> The maximum number of batches will be retained in memory files.</p> <p>Default: 2</p> <p>Maximum count of versions implementation should</p> <p>The value adjusts a trade-off between usage vs cache miss:</p> <ul style="list-style-type: none"> <li>• 2 covers both success cases</li> <li>• 1 covers only success cases</li> <li>• 0 or negative values maximize memory</li> </ul> <p>Used exclusively when HDFSBackedStateStorePlugin.initialize.</p>
spark.sql.streaming.metricsEnabled	<p>Flag whether Dropwizard metrics are reported for active streams.</p> <p>Default: false</p> <p>Use <a href="#">SQLConf.streamMetricsEnabled</a> current value</p>
spark.sql.streaming.minBatchesToRetain	<p><b>(internal)</b> The minimum number of batches for failure recovery</p> <p>Default: 100</p> <p>Use <a href="#">SQLConf.minBatchesToRetain</a> current value</p>

	<p><b>Global watermark policy</b> calculate the global watermark across multiple watermark query</p> <p>Default: <code>min</code></p> <p>Supported values:</p> <ul style="list-style-type: none"> <li>• <code>min</code> - chooses the minimum watermark reported across multiple operators</li> <li>• <code>max</code> - chooses the maximum watermark reported across multiple operators</li> </ul> <p>Cannot be changed before the same checkpoint location.</p>
<code>spark.sql.streaming.multipleWatermarkPolicy</code>	<p>Flag to control whether <code>engine</code> should execute <code>process</code> for eager state streaming queries ( <code>true</code> by default).</p> <p>Default: <code>true</code></p> <p>Use <code>SQLConf.streamingNoDataMicroBatchesEnabled</code> to get the current value.</p>
<code>spark.sql.streaming.noDataProgressEventInterval</code>	<p><b>(internal)</b> How long to wait for events when there is no data. <code>ProgressReporter</code> is responsible for reporting progress.</p> <p>Default: <code>10000L</code></p> <p>Use <code>SQLConf.streamingNoDataProgressEventInterval</code> to get the current value.</p>
<code>spark.sql.streaming.numRecentProgressUpdates</code>	<p>Number of <code>StreamingContext</code> <code>progressBuffer</code> internal <code>ProgressReporter</code> is responsible for streaming query.</p> <p>Default: <code>100</code></p> <p>Use <code>SQLConf.streamingNumRecentProgressUpdates</code> to get the current value.</p>
<code>spark.sql.streaming.pollingDelay</code>	<p><b>(internal)</b> How long (in milliseconds) to wait for StreamExecution before <code>no data was available</code> is triggered.</p> <p>Default: <code>10</code> (milliseconds)</p>

<code>spark.sql.streaming.stateStore.maintenanceInterval</code>	The initial delay and how often the StateStore's <a href="#">maintenance</a> task runs.  Default: 60s
<code>spark.sql.streaming.stateStore.minDeltasForSnapshot</code>	(internal) Minimum number of data files that need to be generated by the HDFSBackedStateStore before performing a snapshot (consolidate).  Default: 10  Use <a href="#">SQLConf.stateStoreMinDeltasForSnapshot</a> to get the current value.
<code>spark.sql.streaming.stateStore.providerClass</code>	(internal) The fully-qualified class name of the <a href="#">StateStoreProvider</a> implementation that manages state data in stateful streams. The provider must have a zero-arg constructor.  Default: <a href="#">HDFSBackedStateStore</a>  Use <a href="#">SQLConf.stateStoreProviderClass</a> to get the current value.
<code>spark.sql.streaming.unsupportedOperationCheck</code>	(internal) When enabled, the <a href="#">StreamingQueryManager</a> will check the <a href="#">plan of a streaming query</a> for unsupported operations only.  Default: true

# StreamingQueryListener — Intercepting Life Cycle Events of Streaming Queries

`StreamingQueryListener` is the [contract](#) of listeners that want to be notified about the [life cycle events](#) of streaming queries, i.e. [start](#), [progress](#) and [termination](#).

Table 1. StreamingQueryListener Contract

Method	Description
<code>onQueryStarted</code>	<pre>onQueryStarted(     event: QueryStartedEvent): Unit</pre> <p>Informs that <code>DataStreamWriter</code> was requested to <a href="#">start execution of the streaming query</a> (on the <a href="#">stream execution thread</a>)</p>
<code>onQueryProgress</code>	<pre>onQueryProgress(     event: QueryProgressEvent): Unit</pre> <p>Informs that <code>MicroBatchExecution</code> has finished <a href="#">triggerExecution phase</a> (the end of a streaming batch)</p>
<code>onQueryTerminated</code>	<pre>onQueryTerminated(     event: QueryTerminatedEvent): Unit</pre> <p>Informs that a streaming query was <a href="#">stopped</a> or terminated due to an error</p>

`StreamingQueryListener` is informed about the [life cycle events](#) when `StreamingQueryListenerBus` is requested to [doPostEvent](#).

Table 2. StreamingQueryListener's Life Cycle Events and Callbacks

Event	Callback	Description
QueryStartedEvent <ul style="list-style-type: none"><li>• <code>id</code></li><li>• <code>runId</code></li><li>• <code>name</code></li></ul>	<code>onQueryStarted</code>	Posted when <code>StreamExecution</code> is requested to run stream processing (when <code>DataStreamWriter</code> is requested to start execution of the streaming query on the stream execution thread)
QueryProgressEvent <ul style="list-style-type: none"><li>• <code>StreamingQueryProgress</code></li></ul>	<code>onQueryProgress</code>	Posted when <code>ProgressReporter</code> is requested to update progress of a streaming query (after <code>MicroBatchExecution</code> has finished triggerExecution phase at the end of a streaming batch)
QueryTerminatedEvent <ul style="list-style-type: none"><li>• <code>id</code></li><li>• <code>runId</code></li><li>• <code>exception</code> if terminated due to an error</li></ul>	<code>onQueryTerminated</code>	Posted when <code>StreamExecution</code> is requested to run stream processing (and the streaming query was stopped or terminated due to an error)

You can register a `StreamingQueryListener` using `StreamingQueryManager.addListener` method.

```
val queryListener: StreamingQueryListener = ...
spark.streams.addListener(queryListener)
```

You can remove a `StreamingQueryListener` using `StreamingQueryManager.removeListener` method.

```
val queryListener: StreamingQueryListener = ...
spark.streams.removeListener(queryListener)
```

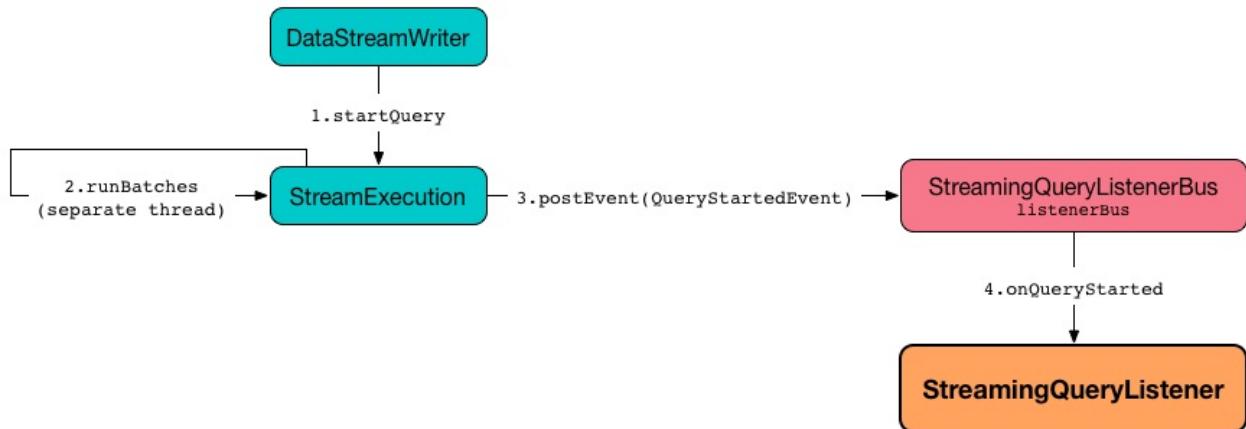


Figure 1. StreamingQueryListener Notified about Query's Start (onQueryStarted)

Note	<code>onQueryStarted</code> is used internally to unblock the <a href="#">starting thread</a> of <code>StreamExecution</code> .
------	---

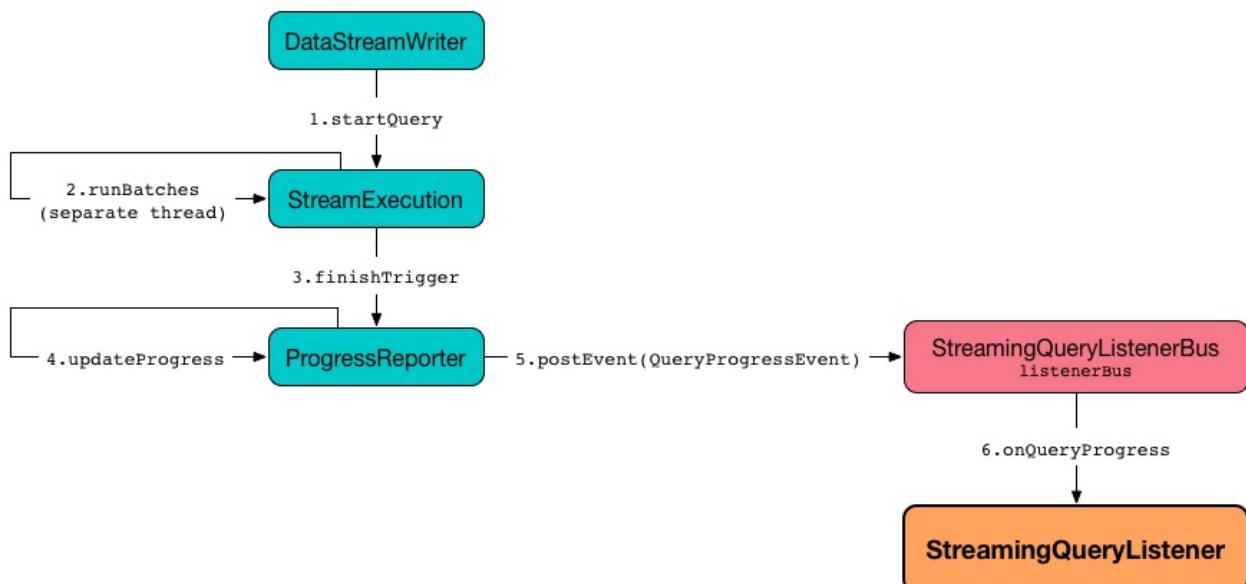


Figure 2. StreamingQueryListener Notified about Query's Progress (onQueryProgress)

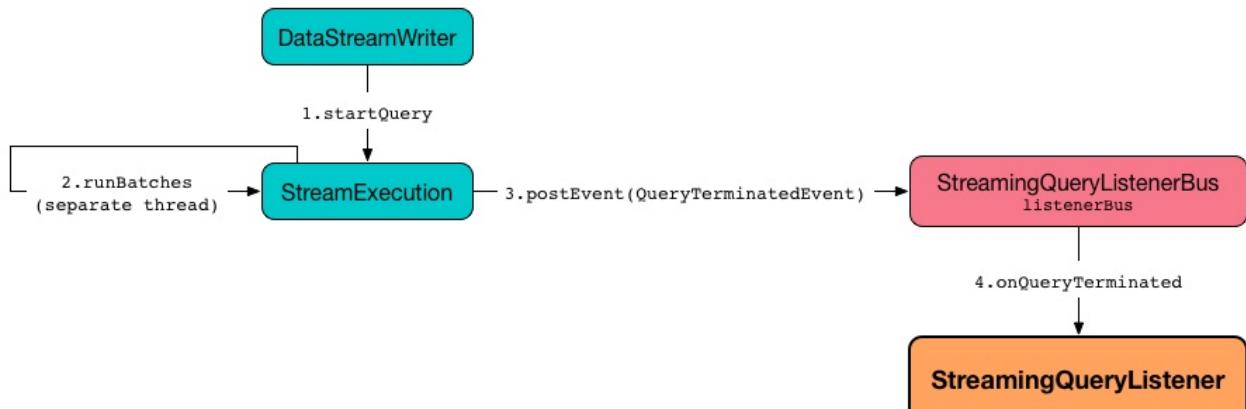


Figure 3. StreamingQueryListener Notified about Query's Termination (onQueryTerminated)

Note	<p>You can also register a streaming event listener using the general <code>SparkListener</code> interface.</p> <p>Read up on <a href="#">SparkListener</a> in the <a href="#">The Internals of Apache Spark</a> book.</p>
------	--



# ProgressReporter Contract

`ProgressReporter` is the [contract](#) of stream execution progress reporters that report the statistics of execution of a streaming query.

Table 1. ProgressReporter Contract

Method	Description
<code>currentBatchId</code>	<pre>currentBatchId: Long</pre> <p><code>Id of the current streaming micro-batch</code></p>
<code>id</code>	<pre>id: UUID</pre> <p><code>Universally unique identifier (UUID) of the streaming query (that stays unchanged between restarts)</code></p>
<code>lastExecution</code>	<pre>lastExecution: QueryExecution</pre> <p><code>QueryExecution of the streaming query</code></p>
<code>logicalPlan</code>	<pre>logicalPlan: LogicalPlan</pre> <p><code>Logical query plan of the streaming query</code></p> <p>Used when <code>ProgressReporter</code> is requested for the following:</p> <ul style="list-style-type: none"> <li>• <a href="#">extract statistics from the most recent query execution</a> (to add <code>watermark</code> metric when a <a href="#">streaming watermark</a> is used)</li> <li>• <a href="#">extractSourceToNumInputRows</a></li> </ul>
<code>name</code>	<pre>name: String</pre> <p><code>Name of the streaming query</code></p>
<code>newData</code>	<pre>newData: Map[BaseStreamingSource, LogicalPlan]</pre> <p><code>Streaming readers and sources with the new data (as a LogicalPlan )</code></p>

	<p>Used when:</p> <ul style="list-style-type: none"> <li>• ProgressReporter <a href="#">extracts statistics from the most recent query execution</a> (to calculate the so-called <code>inputRows</code> )</li> </ul>
<code>offsetSeqMetadata</code>	<pre>offsetSeqMetadata: OffsetSeqMetadata</pre> <p><a href="#">OffsetSeqMetadata</a> (with the current micro-batch <a href="#">event-time watermark and timestamp</a>)</p>
<code>postEvent</code>	<pre>postEvent(event: StreamingQueryListener.Event): Unit</pre> <p>Posts <a href="#">StreamingQueryListener.Event</a></p>
<code>runId</code>	<pre>runId: UUID</pre> <p><a href="#">Universally unique identifier (UUID)</a> of the single run of the streaming query (that changes every restart)</p>
<code>sink</code>	<pre>sink: BaseStreamingSink</pre> <p>The one and only <a href="#">streaming writer or sink</a> of the streaming query</p>
<code>sources</code>	<pre>sources: Seq[BaseStreamingSource]</pre> <p><a href="#">Streaming readers and sources</a> of the streaming query</p> <p>Used when <a href="#">finishing a trigger</a> (and updating progress and <a href="#">marking current status as trigger inactive</a>)</p>
<code>sparkSession</code>	<pre>sparkSession: SparkSession</pre> <p><a href="#">SparkSession</a> of the streaming query</p> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 10px;"> <span style="border: 1px solid #ccc; padding: 2px;">Tip</span> <a href="#">Read up on SparkSession in The Internals of Spark SQL book.</a> </div>
<code>triggerClock</code>	<pre>triggerClock: Clock</pre> <p><a href="#">Clock</a> of the streaming query</p>

Note

[StreamExecution](#) is the one and only known direct extension of the [ProgressReporter Contract](#) in Spark Structured Streaming.

`ProgressReporter` uses the `spark.sql.streaming.noDataProgressEventInterval` configuration property to control how long to wait between two progress events when there is no data (default: `10000L`) when [finishing trigger](#).

`ProgressReporter` uses **yyyy-MM-dd'T'HH:mm:ss.SSS'Z'** time format (with **UTC** timezone).

```

import org.apache.spark.sql.streaming.Trigger
import scala.concurrent.duration._
val sampleQuery = spark
  .readStream
  .format("rate")
  .load
  .writeStream
  .format("console")
  .option("truncate", false)
  .trigger(Trigger.ProcessingTime(10.seconds))
  .start

// Using public API
import org.apache.spark.sql.streaming.SourceProgress
scala> sampleQuery.
|   lastProgress.
|   sources.
|   map { case sp: SourceProgress =>
|     s"source = ${sp.description} => endOffset = ${sp.endOffset}" }.
|   foreach(println)
source = RateSource[rowsPerSecond=1, rampUpTimeSeconds=0, numPartitions=8] => endOffset
t = 663

scala> println(sampleQuery.lastProgress.sources(0))
res40: org.apache.spark.sql.streaming.SourceProgress =
{
  "description" : "RateSource[rowsPerSecond=1, rampUpTimeSeconds=0, numPartitions=8]",
  "startOffset" : 333,
  "endOffset" : 343,
  "numInputRows" : 10,
  "inputRowsPerSecond" : 0.9998000399920015,
  "processedRowsPerSecond" : 200.0
}

// With a hack
import org.apache.spark.sql.execution.streaming.StreamingQueryWrapper
val offsets = sampleQuery.
  asInstanceOf[StreamingQueryWrapper].
  streamingQuery.
  availableOffsets.
  map { case (source, offset) =>
    s"source = $source => offset = $offset" }
scala> offsets.foreach(println)
source = RateSource[rowsPerSecond=1, rampUpTimeSeconds=0, numPartitions=8] => offset =
293

```

Configure logging of the [concrete stream execution progress reporters](#) to see what happens inside a `ProgressReporter`:

**Tip**

- [ContinuousExecution](#)
- [MicroBatchExecution](#)

## **progressBuffer Internal Property**

```
progressBuffer: Queue[StreamingQueryProgress]
```

`progressBuffer` is a [scala.collection.mutable.Queue](#) of [StreamingQueryProgresses](#).

`progressBuffer` has a new `StreamingQueryProgress` added when `ProgressReporter` is requested to [update progress of a streaming query](#).

When the size (the number of `StreamingQueryProgresses`) is above [spark.sql.streaming.numRecentProgressUpdates](#) threshold, the oldest `StreamingQueryProgress` is removed ([dequeued](#)).

`progressBuffer` is used when `ProgressReporter` is requested for the [last](#) and the [recent StreamingQueryProgresses](#)

## **status Method**

```
status: StreamingQueryStatus
```

`status` gives the [current StreamingQueryStatus](#).

**Note**

`status` is used when `StreamingQueryWrapper` is requested for the current status of a streaming query (that is part of [StreamingQuery Contract](#)).

## **Updating Progress of Streaming Query**

### **— updateProgress Internal Method**

```
updateProgress(newProgress: StreamingQueryProgress): Unit
```

`updateProgress` records the input `newProgress` and posts a [QueryProgressEvent](#) event.

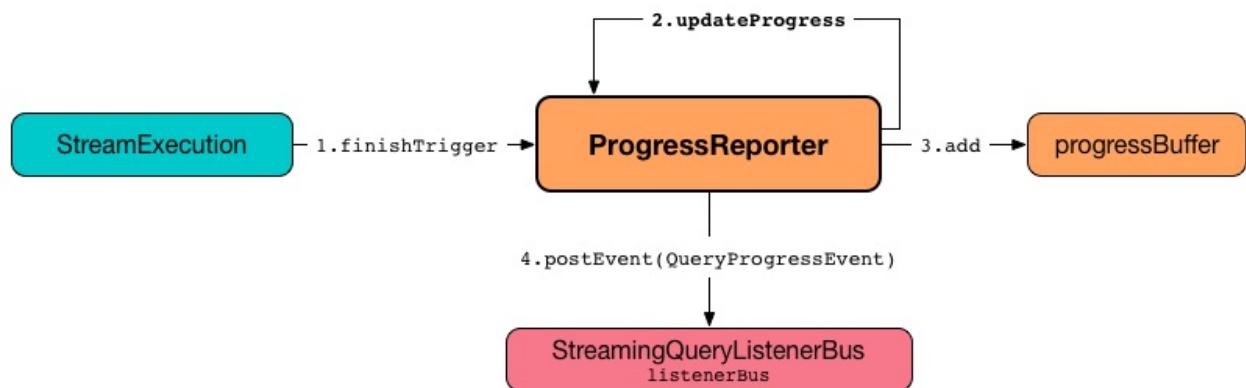


Figure 1. ProgressReporter's Reporting Query Progress

`updateProgress` adds the input `newProgress` to `progressBuffer`.

`updateProgress` removes elements from `progressBuffer` if their number is or exceeds the value of `spark.sql.streaming.numRecentProgressUpdates` property.

`updateProgress` posts a `QueryProgressEvent` (with the input `newProgress` ).

`updateProgress` prints out the following INFO message to the logs:

```
Streaming query made progress: [newProgress]
```

Note	<code>updateProgress</code> synchronizes concurrent access to the <code>progressBuffer</code> internal registry.
Note	<code>updateProgress</code> is used exclusively when <code>ProgressReporter</code> is requested to finish up a trigger.

## Initializing Query Progress for New Trigger — `startTrigger` Method

```
startTrigger(): Unit
```

`startTrigger` prints out the following DEBUG message to the logs:

```
Starting Trigger Calculation
```

Table 2. startTrigger's Internal Registry Changes For New Trigger

Registry	New Value
lastTriggerStartTimestamp	currentTriggerStartTimestamp
currentTriggerStartTimestamp	Requests the trigger clock for the current timestamp (in millis)
currentStatus	Enables ( true ) the isTriggerActive flag of the currentStatus
currentTriggerStartOffsets	null
currentTriggerEndOffsets	null
currentDurationsMs	Clears the currentDurationsMs

Note	<p><code>startTrigger</code> is used when:</p> <ul style="list-style-type: none"> <li>• <code>MicroBatchExecution</code> stream execution engine is requested to run an activated streaming query (at the beginning of every trigger)</li> <li>• <code>ContinuousExecution</code> stream execution engine is requested to run an activated streaming query (at the beginning of every trigger)</li> </ul> <p><code>StreamExecution</code> starts running batches (as part of <code>TriggerExecutor</code> executing a batch runner).</p>

## Finishing Up Streaming Batch (Trigger) and Generating StreamingQueryProgress — `finishTrigger` Method

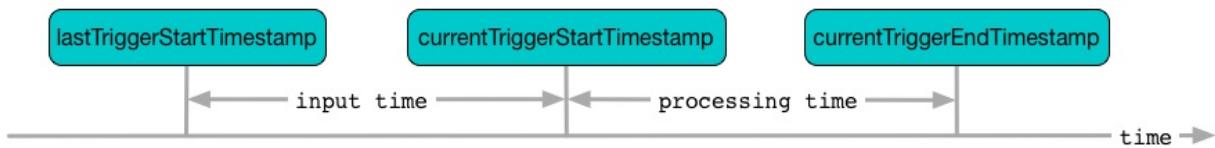
```
finishTrigger(hasNewData: Boolean): Unit
```

Internally, `finishTrigger` sets `currentTriggerEndTimestamp` to the current time (using `triggerClock`).

`finishTrigger` extractExecutionStats.

`finishTrigger` calculates the **processing time** (in seconds) as the difference between the end and `start` timestamps.

`finishTrigger` calculates the **input time** (in seconds) as the difference between the start time of the `current` and `last` triggers.

Figure 2. ProgressReporter's `finishTrigger` and Timestamps

`finishTrigger` prints out the following DEBUG message to the logs:

```
Execution stats: [executionStats]
```

`finishTrigger` creates a [SourceProgress](#) (aka source statistics) for [every source used](#).

`finishTrigger` creates a [SinkProgress](#) (aka sink statistics) for the [sink](#).

`finishTrigger` creates a [StreamingQueryProgress](#).

If there was any data (using the input `hasNewData` flag), `finishTrigger` resets `lastNoDataProgressEventTime` (i.e. becomes the minimum possible time) and updates query progress.

Otherwise, when no data was available (using the input `hasNewData` flag), `finishTrigger` updates query progress only when `lastNoDataProgressEventTime` passed.

In the end, `finishTrigger` disables `isTriggerActive` flag of [StreamingQueryStatus](#) (i.e. sets it to `false`).

#### Note

`finishTrigger` is used exclusively when `MicroBatchExecution` is requested to [run the activated streaming query](#) (after [triggerExecution Phase](#) at the end of a streaming batch).

## Time-Tracking Section (Recording Execution Time for Progress Reporting) — `reportTimeTaken` Method

```
reportTimeTaken[T](
  triggerDetailKey: String)(
  body: => T): T
```

`reportTimeTaken` measures the time to execute `body` and records it in the `currentDurationsMs` internal registry under `triggerDetailKey` key. If the `triggerDetailKey` key was recorded already, the current execution time is added.

In the end, `reportTimeTaken` prints out the following DEBUG message to the logs and returns the result of executing `body`.

```
[triggerDetailKey] took [time] ms
```

`reportTimeTaken` is used when the `stream execution engines` are requested to execute the following phases (that appear as `triggerDetailKey` in the DEBUG message in the logs):

- `MicroBatchExecution`
  - `triggerExecution`
  - `getOffset`
  - `setOffsetRange`
  - `getEndOffset`
  - `walCommit`
  - `getBatch`
  - `queryPlanning`
  - `addBatch`
- `ContinuousExecution`
  - `queryPlanning`
  - `runContinuous`

Note

## Updating Status Message — `updateStatusMessage` Method

```
updateStatusMessage(message: String): Unit
```

`updateStatusMessage` simply updates the `message` in the `StreamingQueryStatus` internal registry.

Note

`updateStatusMessage` is used when:

- `StreamExecution` is requested to run stream processing
- `MicroBatchExecution` is requested to run an activated streaming query, construct the next streaming micro-batch

## Generating Execution Statistics — `extractExecutionStats` Internal Method

```
extractExecutionStats(hasNewData: Boolean): ExecutionStats
```

`extractExecutionStats` generates an [ExecutionStats](#) of the [last execution](#) of the streaming query.

Internally, `extractExecutionStats` generate **watermark** metric (using the [event-time watermark](#) of the [OffsetSeqMetadata](#)) if there is a [EventTimeWatermark](#) unary logical operator in the [logical plan](#) of the streaming query.

**Note**

`EventTimeWatermark` unary logical operator represents [Dataset.withWatermark](#) operator in a streaming query.

`extractExecutionStats` [extractStateOperatorMetrics](#).

`extractExecutionStats` [extractSourceToNumInputRows](#).

`extractExecutionStats` finds the [EventTimeWatermarkExec](#) unary physical operator (with non-zero [EventTimeStats](#)) and generates **max**, **min**, and **avg** statistics.

In the end, `extractExecutionStats` creates a [ExecutionStats](#) with the execution statistics.

If the input `hasNewData` flag is turned off (`false`), `extractExecutionStats` returns an [ExecutionStats](#) with no input rows and event-time statistics (that require data to be processed to have any sense).

**Note**

`extractExecutionStats` is used exclusively when `ProgressReporter` is requested to [finish up a streaming batch \(trigger\)](#) and generate a [StreamingQueryProgress](#).

## Generating StateStoreWriter Metrics ([StateOperatorProgress](#))

### — **extractStateOperatorMetrics Internal Method**

```
extractStateOperatorMetrics(  
    hasNewData: Boolean): Seq[StateOperatorProgress]
```

`extractStateOperatorMetrics` requests the [QueryExecution](#) for the optimized execution plan (`executedPlan`) and finds all [StateStoreWriter](#) physical operators and requests them for [StateOperatorProgress](#).

`extractStateOperatorMetrics` clears (zeros) the **numRowsUpdated** metric for the given `hasNewData` turned off (`false`).

`extractStateOperatorMetrics` returns an empty collection for the [QueryExecution](#) uninitialized (`null`).

**Note**

`extractStateOperatorMetrics` is used exclusively when `ProgressReporter` is requested to [generate execution statistics](#).

## extractSourceToNumInputRows Internal Method

```
extractSourceToNumInputRows(): Map[BaseStreamingSource, Long]
```

`extractSourceToNumInputRows` ...FIXME

**Note**

`extractSourceToNumInputRows` is used exclusively when `ProgressReporter` is requested to [generate execution statistics](#).

## formatTimestamp Internal Method

```
formatTimestamp(millis: Long): String
```

`formatTimestamp` ...FIXME

**Note**

`formatTimestamp` is used when...FIXME

## Recording Trigger Offsets (StreamProgress) — recordTriggerOffsets Method

```
recordTriggerOffsets(  
  from: StreamProgress,  
  to: StreamProgress): Unit
```

`recordTriggerOffsets` simply sets (*records*) the `currentTriggerStartOffsets` and `currentTriggerEndOffsets` internal registries to the `json` representations of the `from` and `to` `StreamProgresses`.

**Note**

`recordTriggerOffsets` is used when:

- `MicroBatchExecution` is requested to [run the activated streaming query](#)
- `ContinuousExecution` is requested to [commit an epoch](#)

## Last StreamingQueryProgress — lastProgress Method

```
lastProgress: StreamingQueryProgress
```

`lastProgress` ...FIXME

Note

`lastProgress` is used when...FIXME

## recentProgress Method

```
recentProgress: Array[StreamingQueryProgress]
```

`recentProgress` ...FIXME

Note

`recentProgress` is used when...FIXME

## Internal Properties

Name	Description
<code>currentDurationsMs</code>	<p><code>scala.collection.mutable.HashMap</code> of action names (aka <code>triggerDetailKey</code>) and their cumulative times (in milliseconds).</p> <p>Starts empty when <code>ProgressReporter</code> sets the state for a batch with new entries added or updated when reporting execution time (of an action).</p> <p><b>Tip</b></p> <p>You can see the current value of <code>currentDurationsMs</code> in progress reports under <code>durationMs</code>.</p> <pre>scala&gt; query.lastProgress.durationMs res3: java.util.Map[String,Long] = {triggerExecution=60, queryPlanning=1, getBatch=5, getOffset=0, addBatch=30, walCommit=23}</pre>
<code>currentStatus</code>	<p><code>StreamingQueryStatus</code> with the current status of the streaming query</p> <p>Available using <code>status</code> method</p> <ul style="list-style-type: none"> <li>message updated with <code>updateStatusMessage</code></li> </ul>
<code>currentTriggerEndOffsets</code>	Timestamp of when the current batch/trigger has ended

	<p>Default: <code>-1L</code></p>
<code>currentTriggerStartOffsets</code>	<p><code>currentTriggerStartOffsets: Map[BaseStreamingSource, :]</code></p> <p>Start offsets (in <a href="#">JSON format</a>) per source</p> <p>Used exclusively when <a href="#">finishing up a streaming batch</a> (triggering and generating <a href="#">StreamingQueryProgress</a> (for a <a href="#">SourceProvider</a>))</p> <p>Reset (<code>null</code>) when <a href="#">initializing a query progress</a> for a new trigger</p> <p>Initialized when <a href="#">recording trigger offsets</a> (<a href="#">StreamProgress</a>)</p>
<code>currentTriggerStartTimestamp</code>	<p>Timestamp of when the current batch/trigger has started</p> <p>Default: <code>-1L</code></p>
<code>lastNoDataProgressEventTime</code>	<p>Default: <code>Long.MinValue</code></p>
<code>lastTriggerStartTimestamp</code>	<p>Timestamp of when the last batch/trigger started</p> <p>Default: <code>-1L</code></p>
<code>metricWarningLogged</code>	<p>Flag to...FIXME</p> <p>Default: <code>false</code></p>

# StreamingQueryProgress

`StreamingQueryProgress` holds information about the progress of a streaming query.

`StreamingQueryProgress` is created exclusively when `StreamExecution` finishes a trigger.

Note

Use `lastProgress` property of a `StreamingQuery` to access the most recent `StreamingQueryProgress` update.

```
val sq: StreamingQuery = ...
sq.lastProgress
```

Note

Use `recentProgress` property of a `StreamingQuery` to access the most recent `StreamingQueryProgress` updates.

```
val sq: StreamingQuery = ...
sq.recentProgress
```

Note

Use `StreamingQueryListener` to get notified about `StreamingQueryProgress` updates while a streaming query is executed.

Table 1. StreamingQueryProgress's Properties

Name	Description
id	Unique identifier of a streaming query
runId	Unique identifier of the current execution of a streaming query
name	Optional query name
timestamp	Time when the trigger has started (in ISO8601 format).
batchId	Unique id of the current batch
durationMs	Durations of the internal phases (in milliseconds)
eventTime	Statistics of event time seen in this batch
stateOperators	Information about stateful operators in the query that store state.
sources	Statistics about the data read from every streaming source in a streaming query
sink	Information about progress made for a sink

# ExecutionStats

ExecutionStats is...FIXME

# SourceProgress

SourceProgress is...FIXME

# SinkProgress

SinkProgress is...FIXME

# StreamingQueryStatus

StreamingQueryStatus is...FIXME

# MetricsReporter

MetricsReporter is...FIXME

# Web UI

Web UI...FIXME

Caution	FIXME What's visible on the plan diagram in the SQL tab of the UI
---------	---

# Logging

Caution	FIXME
---------	-------

# FileStreamSource

`FileStreamSource` is a [Source](#) that reads text files from `path` directory as they appear. It uses `LongOffset` offsets.

Note

It is used by [DataSource.createSource](#) for `FileFormat`.

You can provide the `schema` of the data and `dataFrameBuilder` - the function to build a `DataFrame` in [getBatch](#) at instantiation time.

```
// NOTE The source directory must exist
// mkdir text-logs

val df = spark.readStream
  .format("text")
  .option("maxFilesPerTrigger", 1)
  .load("text-logs")

scala> df.printSchema
root
 |-- value: string (nullable = true)
```

Batches are indexed.

It lives in `org.apache.spark.sql.execution.streaming` package.

```
import org.apache.spark.sql.types._
val schema = StructType(
  StructField("id", LongType, nullable = false) ::
  StructField("name", StringType, nullable = false) ::
  StructField("score", DoubleType, nullable = false) :: Nil)

// You should have input-json directory available
val in = spark.readStream
  .format("json")
  .schema(schema)
  .load("input-json")

val input = in.transform { ds =>
  println("transform executed") // <-- it's going to be executed once only
  ds
}

scala> input.isStreaming
res9: Boolean = true
```

It tracks already-processed files in `seenFiles` hash map.

**Tip** Enable `DEBUG` or `TRACE` logging level for `org.apache.spark.sql.execution.streaming.FileStreamSource` to see what happens inside.

Add the following line to `conf/log4j.properties` :

```
log4j.logger.org.apache.spark.sql.execution.streaming.FileStreamSource=TRACE
```

Refer to [Logging](#).

## Creating FileStreamSource Instance

Caution

FIXME

## Options

### maxFilesPerTrigger

`maxFilesPerTrigger` option specifies the maximum number of files per trigger (batch). It limits the file stream source to read the `maxFilesPerTrigger` number of files specified at a time and hence enables rate limiting.

It allows for a static set of files be used like a stream for testing as the file set is processed `maxFilesPerTrigger` number of files at a time.

### schema

If the schema is specified at instantiation time (using optional `dataSchema` constructor parameter) it is returned.

Otherwise, `fetchAllFiles` internal method is called to list all the files in a directory.

When there is at least one file the schema is calculated using `dataFrameBuilder` constructor parameter function. Else, an `IllegalArgumentException("No schema specified")` is thrown unless it is for `text` provider (as `providerName` constructor parameter) where the default schema with a single `value` column of type `StringType` is assumed.

Note

**text** as the value of `providerName` constructor parameter denotes **text file stream provider**.

## getOffset Method

```
getOffset: Option[Offset]
```

**Note**

`getOffset` is part of the [Source Contract](#) to find the latest `offset`.

`getOffset ...FIXME`

The maximum offset (`getOffset`) is calculated by fetching all the files in `path` excluding files that start with `_` (underscore).

When computing the maximum offset using `getOffset`, you should see the following DEBUG message in the logs:

```
Listed ${files.size} in ${(endTime.toDouble - startTime) / 1000000}ms
```

When computing the maximum offset using `getOffset`, it also filters out the files that were already seen (tracked in `seenFiles` internal registry).

You should see the following DEBUG message in the logs (depending on the status of a file):

```
new file: $file
// or
old file: $file
```

## Generating DataFrame for Streaming Batch — `getBatch` Method

`FileStreamSource.getBatch` asks [metadataLog](#) for the batch.

You should see the following INFO and DEBUG messages in the logs:

```
INFO Processing ${files.length} files from ${startId + 1}:$endId
DEBUG Streaming ${files.mkString(", ")}
```

The method to create a result batch is given at instantiation time (as `dataFrameBuilder` constructor parameter).

## metadataLog

`metadataLog` is a metadata storage using `metadataPath` path (which is a constructor parameter).

Note	It extends <code>HDFSMetadataLog[Seq[String]]</code> .
------	--

Caution	<code>FIXME Review</code> <code>HDFSMetadataLog</code>
---------	--

## fetchMaxOffset Internal Method

```
fetchMaxOffset(): FileStreamSourceOffset
```

`fetchMaxOffset` ...`FIXME`

Note	<code>fetchMaxOffset</code> is used exclusively when <code>FileStreamSource</code> is requested to <a href="#">getOffset</a> .
------	--

## fetchAllFiles Internal Method

```
fetchAllFiles(): Seq[(String, Long)]
```

`fetchAllFiles` ...`FIXME`

Note	<code>fetchAllFiles</code> is used exclusively when <code>FileStreamSource</code> is requested to <a href="#">fetchMaxOffset</a> .
------	--

## allFilesUsingMetadataLogFileIndex Internal Method

```
allFilesUsingMetadataLogFileIndex(): Seq[FileStatus]
```

`allFilesUsingMetadataLogFileIndex` simply creates a new [MetadataLogFileIndex](#) and requests it to `allFiles` .

Note	<code>allFilesUsingMetadataLogFileIndex</code> is used exclusively when <code>FileStreamSource</code> is requested to <a href="#">fetchAllFiles</a> (when requested for <a href="#">fetchMaxOffset</a> when <code>FileStreamSource</code> is requested to <a href="#">getOffset</a> ).
------	--

# FileStreamSink — Streaming Sink for File-Based Data Sources

`FileStreamSink` is a concrete [streaming sink](#) that writes out the results of a streaming query to files (of the specified [FileFormat](#)) in the [root path](#).

```
import scala.concurrent.duration._
import org.apache.spark.sql.streaming.{OutputMode, Trigger}
val sq = in.
  writeStream.
  format("parquet").
  option("path", "parquet-output-dir").
  option("checkpointLocation", "checkpoint-dir").
  trigger(Trigger.ProcessingTime(10.seconds)).
  outputMode(OutputMode.Append).
  start
```

`FileStreamSink` is [created](#) exclusively when `DataSource` is requested to [create a streaming sink](#) for a file-based data source (i.e. `FileFormat` ).

Tip	Read up on <a href="#">FileFormat</a> in <a href="#">The Internals of Spark SQL</a> book.
-----	---

`FileStreamSink` supports [Append output mode](#) only.

`FileStreamSink` uses `spark.sql.streaming.fileSink.log.deletion` (as `isDeletingExpiredLog` )

The textual representation of `FileStreamSink` is **FileSink[path]**

`FileStreamSink` uses **\_spark\_metadata** directory for...FIXME

Tip	<p>Enable <code>ALL</code> logging level for <code>org.apache.spark.sql.execution.streaming.FileStreamSink</code> to see what happens inside.</p> <p>Add the following line to <code>conf/log4j.properties</code> :</p> <pre>log4j.logger.org.apache.spark.sql.execution.streaming.FileStreamSink=ALL</pre> <p>Refer to <a href="#">Logging</a>.</p>
-----	--

## Creating FileStreamSink Instance

`FileStreamSink` takes the following to be created:

- `SparkSession`
- Root directory
- `FileFormat`
- Names of the partition columns
- Configuration options

`FileStreamSink` initializes the [internal properties](#).

## "Adding" Batch of Data to Sink — `addBatch` Method

```
addBatch(  
    batchId: Long,  
    data: DataFrame): Unit
```

Note	<code>addBatch</code> is a part of <a href="#">Sink Contract</a> to "add" a batch of data to the sink.
------	--

`addBatch` ...FIXME

## Creating BasicWriteJobStatsTracker — `basicWriteJobStatsTracker` Internal Method

```
basicWriteJobStatsTracker: BasicWriteJobStatsTracker
```

`basicWriteJobStatsTracker` simply creates a `BasicWriteJobStatsTracker` with the basic metrics:

- number of written files
- bytes of written output
- number of output rows
- number of dynamic partitions

Tip	Read up on <a href="#">BasicWriteJobStatsTracker</a> in <a href="#">The Internals of Spark SQL</a> book.
-----	--

Note	<code>basicWriteJobStatsTracker</code> is used exclusively when <code>FileStreamSink</code> is requested to <a href="#">addBatch</a> .
------	--

## `hasMetadata` Object Method

```
hasMetadata(  
    path: Seq[String],  
    hadoopConf: Configuration): Boolean
```

hasMetadata ...FIXME

Note	<p><code>hasMetadata</code> is used when:</p> <ul style="list-style-type: none"> <li>• <code>DataSource</code> (Spark SQL) is requested to resolve a <code>FileFormat</code> relation (<code>resolveRelation</code>) and creates a <code>HadoopFsRelation</code></li> <li>• <code>FileStreamSource</code> is requested to <code>fetchAllFiles</code></li> </ul>
------	---

## Internal Properties

Name	Description
<code>basePath</code>	<p><b>Base path</b> (Hadoop's <code>Path</code> for the given <code>path</code>)  Used when...FIXME</p>
<code>logPath</code>	<p><b>Metadata log path</b> (Hadoop's <code>Path</code> for the <code>base path</code> and the <code>_spark_metadata</code>)  Used exclusively to create the <code>FileStreamSinkLog</code></p>
<code>fileLog</code>	<p><code>FileStreamSinkLog</code> (for the <code>version 1</code> and the <code>metadata log path</code>)  Used exclusively when <code>FileStreamSink</code> is requested to <code>addBatch</code></p>
<code>hadoopConf</code>	<p>Hadoop's <code>Configuration</code>  Used when...FIXME</p>

# FileStreamSinkLog

`FileStreamSinkLog` is a concrete [CompactibleFileStreamLog](#) (of [SinkFileStatuses](#)) for [FileStreamSink](#) and [MetadataLogFileIndex](#).

`FileStreamSinkLog` uses **1** for the version.

`FileStreamSinkLog` uses **add** action to create new [metadata logs](#).

`FileStreamSinkLog` uses **delete** action to mark [metadata logs](#) that should be excluded from [compaction](#).

## Creating FileStreamSinkLog Instance

`FileStreamSinkLog` (like the parent [CompactibleFileStreamLog](#)) takes the following to be created:

- Metadata version
- `SparkSession`
- Path of the metadata log directory

### compactLogs Method

```
compactLogs(logs: Seq[SinkFileStatus]): Seq[SinkFileStatus]
```

Note	<code>compactLogs</code> is part of the <a href="#">CompactibleFileStreamLog Contract</a> to...FIXME.
------	---

`compactLogs` ...FIXME

# SinkFileStatus

`SinkFileStatus` represents the status of files of [FileStreamSink](#) (and the type of the metadata of [FileStreamSinkLog](#)):

- Path
- Size
- `isDir` flag
- Modification time
- Block replication
- Block size
- Action (either [add](#) or [delete](#))

## toFileStatus Method

```
toFileStatus: FileStatus
```

`toFileStatus` simply creates a new Hadoop [FileStatus](#).

Note	<code>toFileStatus</code> is used exclusively when <code>MetadataLogFileIndex</code> is <a href="#">created</a> .
------	---

## Creating SinkFileStatus Instance — apply Object Method

```
apply(f: FileStatus): SinkFileStatus
```

`apply` simply creates a new [SinkFileStatus](#) (with [add](#) action).

Note	<code>apply</code> is used exclusively when <code>ManifestFileCommitProtocol</code> is requested to <a href="#">commitTask</a> .
------	--

# ManifestFileCommitProtocol

`ManifestFileCommitProtocol` is...FIXME

## commitJob Method

```
commitJob(  
    jobContext: JobContext,  
    taskCommits: Seq[TaskCommitMessage]): Unit
```

Note `commitJob` is part of the `FileCommitProtocol` contract to...FIXME.

`commitJob` ...FIXME

## commitTask Method

```
commitTask(  
    taskContext: TaskAttemptContext): TaskCommitMessage
```

Note `commitTask` is part of the `FileCommitProtocol` contract to...FIXME.

`commitTask` ...FIXME

# MetadataLogFileIndex

`MetadataLogFileIndex` is a `PartitioningAwareFileIndex` of [metadata log files](#) (generated by [FileStreamSink](#)).

`MetadataLogFileIndex` is [created](#) when:

- `DataSource` (Spark SQL) is requested to resolve a `FileFormat` relation (`resolveRelation`) and creates a `HadoopFsRelation`
- `FileStreamSource` is requested to `allFilesUsingMetadataLogFileIndex`

<span style="font-size: 1.2em;">Tip</span>	<p>Enable <code>ALL</code> logging level for  <code>org.apache.spark.sql.execution.streaming.MetadataLogFileIndex</code> to see what happens inside.</p> <p>Add the following line to <code>conf/log4j.properties</code> :</p> <pre>log4j.logger.org.apache.spark.sql.execution.streaming.MetadataLogFileIndex=ALL</pre> <p>Refer to <a href="#">Logging</a>.</p>
--	---

## Creating MetadataLogFileIndex Instance

`MetadataLogFileIndex` takes the following to be created:

- `SparkSession`
- Hadoop's [Path](#)
- User-defined schema (`Option[StructType]`)

`MetadataLogFileIndex` initializes the [internal properties](#).

While being created, `MetadataLogFileIndex` prints out the following INFO message to the logs:

```
Reading streaming file log from [metadataDirectory]
```

## Internal Properties

Name	Description
metadataDirectory	<b>Metadata directory</b> (Hadoop's <a href="#">Path</a> of the <code>_spark_metadata</code> directory under the <a href="#">path</a> ) Used when...FIXME
metadataLog	<a href="#">FileStreamSinkLog</a> (with the <code>_spark_metadata</code> directory)
allFilesFromLog	<b>Metadata log files</b>

# Kafka Data Source — Streaming Data Source for Apache Kafka

**Kafka Data Source** is the streaming data source for [Apache Kafka](#) in Spark Structured Streaming.

Kafka Data Source provides a [streaming source](#) and a [streaming sink](#) for [micro-batch](#) and [continuous](#) stream processing.

## spark-sql-kafka-0-10 External Module

Kafka Data Source is part of the **spark-sql-kafka-0-10** external module that is distributed with the official distribution of Apache Spark, but it is not included in the CLASSPATH by default.

You should define `spark-sql-kafka-0-10` module as part of the build definition in your Spark project, e.g. as a `libraryDependency` in `build.sbt` for sbt:

```
libraryDependencies += "org.apache.spark" %% "spark-sql-kafka-0-10" % "2.4.4"
```

For Spark environments like `spark-submit` (and "derivatives" like `spark-shell`), you should use `--packages` command-line option:

```
./bin/spark-shell --packages org.apache.spark:spark-sql-kafka-0-10_2.12:2.4.4
```

Note	Replace the version of <code>spark-sql-kafka-0-10</code> module (e.g. <code>2.4.4</code> above) with one of the available versions found at <a href="#">The Central Repository's Search</a> that matches your version of Apache Spark.
------	--

## Streaming Source

With [spark-sql-kafka-0-10 module](#) you can use **kafka** data source format for loading data (reading records) from one or more Kafka topics as a streaming Dataset.

```
val records = spark
  .readStream
  .format("kafka")
  .option("subscribePattern", """topic-\d{2}""") // topics with two digits at the end
  .option("kafka.bootstrap.servers", ":9092")
  .load
```

Kafka data source supports many options for reading.

Internally, the **kafka** data source format for reading is available through **KafkaSourceProvider** that is a [MicroBatchReadSupport](#) and [ContinuousReadSupport](#) for [micro-batch](#) and [continuous](#) stream processing, respectively.

## Predefined (Fixed) Schema

Kafka Data Source uses a predefined (fixed) schema.

Table 1. Kafka Data Source's Fixed Schema (in the positional order)

Name	Type
key	BinaryType
value	BinaryType
topic	StringType
partition	IntegerType
offset	LongType
timestamp	TimestampType
timestampType	IntegerType

```
scala> records.printSchema
root
|-- key: binary (nullable = true)
|-- value: binary (nullable = true)
|-- topic: string (nullable = true)
|-- partition: integer (nullable = true)
|-- offset: long (nullable = true)
|-- timestamp: timestamp (nullable = true)
|-- timestampType: integer (nullable = true)
```

Internally, the fixed schema is defined as part of the `DataSourceReader` contract through [MicroBatchReader](#) and [ContinuousReader](#) extension contracts for [micro-batch](#) and [continuous](#) stream processing, respectively.

Tip

Read up on [DataSourceReader](#) in [The Internals of Spark SQL](#) book.

Use `Column.cast` operator to cast `BinaryType` to a `StringType` (for `key` and `value` columns).

```
scala> :type records
org.apache.spark.sql.DataFrame

val values = records
  .select($"value" cast "string") // deserializing values
values.printSchema
root
 |-- value: string (nullable = true)
```

Tip

## Streaming Sink

With [spark-sql-kafka-0-10 module](#) you can use **kafka** data source format for writing the result of executing a streaming query (a streaming Dataset) to one or more Kafka topics.

```
val sq = records
  .writeStream
  .format("kafka")
  .option("kafka.bootstrap.servers", ":9092")
  .option("topic", "kafka2console-output")
  .option("checkpointLocation", "checkpointLocation-kafka2console")
  .start
```

Internally, the **kafka** data source format for writing is available through [KafkaSourceProvider](#) that is a [StreamWriterSupport](#).

## Micro-Batch Stream Processing

Kafka Data Source supports [Micro-Batch Stream Processing](#) (i.e. `Trigger.Once` and `Trigger.ProcessingTime` triggers) via [KafkaMicroBatchReader](#).

```
import org.apache.spark.sql.streaming.Trigger
import scala.concurrent.duration._

val sq = spark
  .readStream
  .format("kafka")
  .option("subscribePattern", "kafka2console.*")
  .option("kafka.bootstrap.servers", ":9092")
  .load
  .withColumn("value", $"value" cast "string") // deserializing values
  .writeStream
  .format("console")
  .option("truncate", false) // format-specific option
  .option("checkpointLocation", "checkpointLocation-kafka2console") // generic query option
  .trigger(Trigger.ProcessingTime(30.seconds))
  .queryName("kafka2console-microbatch")
  .start

// In the end, stop the streaming query
sq.stop
```

Kafka Data Source can assign a single task per Kafka partition (using [KafkaOffsetRangeCalculator](#) in [Micro-Batch Stream Processing](#)).

Kafka Data Source can reuse a Kafka consumer (using [KafkaMicroBatchReader](#) in [Micro-Batch Stream Processing](#)).

## Continuous Stream Processing

Kafka Data Source supports [Continuous Stream Processing](#) (i.e. `Trigger.Continuous` trigger) via [KafkaContinuousReader](#).

```

import org.apache.spark.sql.streaming.Trigger
import scala.concurrent.duration._

val sq = spark
  .readStream
  .format("kafka")
  .option("subscribePattern", "kafka2console.*")
  .option("kafka.bootstrap.servers", ":9092")
  .load
  .withColumn("value", $"value" cast "string") // convert bytes to string for display
purposes
  .writeStream
  .format("console")
  .option("truncate", false) // format-specific option
  .option("checkpointLocation", "checkpointLocation-kafka2console") // generic query o
ption
  .queryName("kafka2console-continuous")
  .trigger(Trigger.Continuous(10.seconds))
  .start

// In the end, stop the streaming query
sq.stop

```

## Configuration Options

Note	Options with <b>kafka.</b> prefix (e.g. <code>kafka.bootstrap.servers</code> ) are considered configuration properties for the Kafka consumers used on the <code>driver</code> and <code>executors</code> .
------	---

Table 2. Kafka Data Source's Options (Case-Insensitive)

Option	Description
<code>assign</code>	<p>Topic subscription strategy that accepts a JSON with topics and partitions, e.g.</p> <pre>{"topicA": [0,1], "topicB": [0,1]}</pre> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 10px;"> <p>Note Exactly one topic subscription strategy is allowed (that <code>KafkaSourceProvider</code> <code>validates</code> before creating <code>KafkaSource</code> ).</p> </div>
<code>failOnDataLoss</code>	<p>Flag to control whether...FIXME</p> <p>Default: <code>true</code></p> <p>Used when <code>KafkaSourceProvider</code> is requested for <code>failOnDataLoss</code> configuration property</p>
<code>kafka.bootstrap.servers</code>	( <b>required</b> ) <code>bootstrap.servers</code> configuration property of the Kafka consumers used on the driver and executors

	<p><b>Default:</b> (empty)</p>
kafkaConsumer.pollTimeoutMs	<p>The time (in milliseconds) spent waiting in <code>consumer.poll</code> is not available in the buffer.</p> <p><b>Default:</b> <code>spark.network.timeout</code> or 120s</p> <p>Used when...FIXME</p>
maxOffsetsPerTrigger	<p>Number of records to fetch per trigger (to limit the number of records to fetch).</p> <p><b>Default:</b> (undefined)</p> <p>Unless defined, <code>KafkaSource</code> requests <a href="#">KafkaOffsetReader latest offsets</a>.</p>
minPartitions	<p>Minimum number of partitions per executor (given Kafka partitions)</p> <p><b>Default:</b> (undefined)</p> <p>Must be undefined (default) or greater than 0</p> <p>When undefined (default) or smaller than the number of <code>TopicPartitions</code> with records to consume from, <code>KafkaMicroBatchReader</code> uses <a href="#">KafkaOffsetRangeCalculator</a> the preferred executor for every <code>TopicPartition</code> (and the available executors).</p>
startingOffsets	<p>Starting offsets</p> <p><b>Default:</b> latest</p> <p>Possible values:</p> <ul style="list-style-type: none"> <li>• latest</li> <li>• earliest</li> <li>• JSON with topics, partitions and their starting offsets,</li> </ul> <pre>{"topicA":{"part":offset,"p1":-1}, "topicB":{"0":-2}}</pre> <p><b>Tip</b> Use Scala's triple quotes for the JSON for topic partitions and offsets.</p> <pre>option(   "startingOffsets",   """{"topic1":{"0":5, "4":-1}, "topic2":{"0":-2}}""" )</pre>

	<p><a href="#">Topic subscription strategy</a> that accepts topic names as a separated string, e.g.</p> <pre>topic1,topic2,topic3</pre>
subscribe	<p><b>Note</b> Exactly one topic subscription strategy is allowed (that <code>kafkaSourceProvider validates</code> before creating <code>KafkaSource</code> ).</p>
	<p><a href="#">Topic subscription strategy</a> that uses Java's <code>java.util.regex</code> for the topic subscription regex pattern of topics to subscribe, e.g.</p> <pre>topic\d</pre>
subscribepattern	<p><b>Tip</b> Use Scala's triple quotes for the regular expression for topic subscription regex pattern.</p> <pre>option("subscribepattern", """topic\d""")</pre>
	<p><b>Note</b> Exactly one topic subscription strategy is allowed (that <code>kafkaSourceProvider validates</code> before creating <code>KafkaSource</code> ).</p>
topic	<p>Optional topic name to use for writing a streaming query Default: (empty)</p> <p>Unless defined, Kafka data source uses the topic names defined in the <code>topic</code> field in the incoming data.</p>

## Logical Query Plan for Reading

When `DataStreamReader` is requested to load a dataset with **kafka** data source format, it creates a DataFrame with a [StreamingRelationV2](#) leaf logical operator.

```
scala> records.explain(extended = true)
== Parsed Logical Plan ==
StreamingRelationV2 org.apache.spark.sql.kafka010.KafkaSourceProvider@1a366d0, kafka,
Map(maxOffsetsPerTrigger -> 1, startingOffsets -> latest, subscribepattern -> topic\d,
kafka.bootstrap.servers -> :9092), [key#7, value#8, topic#9, partition#10, offset#11L
, timestamp#12, timestampType#13], StreamingRelation DataSource(org.apache.spark.sql.S
parkSession@39b3de87,kafka,List(),None,List(),None,Map(maxOffsetsPerTrigger -> 1, star
tingOffsets -> latest, subscribepattern -> topic\d, kafka.bootstrap.servers -> :9092),
None), kafka, [key#0, value#1, topic#2, partition#3, offset#4L, timestamp#5, timestamp
Type#6]
...
...
```

## Logical Query Plan for Writing

When `DataStreamWriter` is requested to start a streaming query with **kafka** data source format for writing, it requests the `StreamingQueryManager` to [create a streaming query](#) that in turn creates (a `StreamingQueryWrapper` with) a [ContinuousExecution](#) or a [MicroBatchExecution](#) for [continuous](#) and [micro-batch](#) stream processing, respectively.

```
scala> sq.explain(extended = true)
== Parsed Logical Plan ==
WriteToDataSourceV2 org.apache.spark.sql.execution.streaming.sources.MicroBatchWriter@
bf98b73
+- Project [key#28 AS key#7, value#29 AS value#8, topic#30 AS topic#9, partition#31 AS
partition#10, offset#32L AS offset#11L, timestamp#33 AS timestamp#12, timestampType#34
AS timestampType#13]
  +- Streaming RelationV2 kafka[key#28, value#29, topic#30, partition#31, offset#32L,
timestamp#33, timestampType#34] (Options: [subscribePattern=kafka2console.*, kafka.boot
strap.servers=:9092])
```

## Demo: Streaming Aggregation with Kafka Data Source

Check out [Demo: Streaming Aggregation with Kafka Data Source](#).

	Use the following to publish events to Kafka.
Tip	<pre>// 1st streaming batch \$ cat /tmp/1 1,1,1 15,2,1  \$ kafkacat -P -b localhost:9092 -t topic1 -l /tmp/1  // Alternatively (and slower due to JVM bootup) \$ cat /tmp/1   ./bin/kafka-console-producer.sh --topic topic1 --broker-list loca</pre>



# KafkaSourceProvider — Data Source Provider for Apache Kafka

`KafkaSourceProvider` is a `DataSourceRegister` and registers a developer-friendly alias for **kafka** data source format in Spark Structured Streaming.

Tip	Read up on <a href="#">DataSourceRegister</a> in <a href="#">The Internals of Spark SQL</a> book.
-----	---

`KafkaSourceProvider` supports [micro-batch stream processing](#) (through `MicroBatchReadSupport` contract) and [creates a specialized KafkaMicroBatchReader](#).

`KafkaSourceProvider` requires the following options (that you can set using `option` method of `DataStreamReader` or `DataStreamWriter`):

- Exactly one of the following options: [subscribe](#), [subscribePattern](#) or [assign](#)
- `kafka.bootstrap.servers`

Tip	Refer to <a href="#">Kafka Data Source's Options</a> for the supported configuration options.
-----	---

Internally, `KafkaSourceProvider` sets the [properties for Kafka Consumers on executors](#) (that are passed on to `InternalKafkaConsumer` when requested to create a Kafka consumer with a single `TopicPartition` manually assigned).

Table 1. KafkaSourceProvider's Properties for Kafka Consumers on Executors

ConsumerConfig's Key	Value	Description
<code>KEY_DESERIALIZER_CLASS_CONFIG</code>	<code>ByteArrayDeserializer</code>	FIXME
<code>VALUE_DESERIALIZER_CLASS_CONFIG</code>	<code>ByteArrayDeserializer</code>	FIXME
<code>AUTO_OFFSET_RESET_CONFIG</code>	<code>none</code>	FIXME
<code>GROUP_ID_CONFIG</code>	<code>uniqueGroupId-executor</code>	FIXME
<code>ENABLE_AUTO_COMMIT_CONFIG</code>	<code>false</code>	FIXME
<code>RECEIVE_BUFFER_CONFIG</code>	<code>65536</code>	Only when not set in the <a href="#">specifiedKafkaParams</a> already

Enable ALL logging levels for `org.apache.spark.sql.kafka010.KafkaSourceProvider` logger to see what happens inside.

Add the following line to `conf/log4j.properties`:

Tip

```
log4j.logger.org.apache.spark.sql.kafka010.KafkaSourceProvider=ALL
```

Refer to [Logging](#).

## Creating Streaming Source — `createSource` Method

```
createSource(  
    sqlContext: SQLContext,  
    metadataPath: String,  
    schema: Option[StructType],  
    providerName: String,  
    parameters: Map[String, String]): Source
```

Note

`createSource` is part of the [StreamSourceProvider Contract](#) to create a [streaming source](#) for a format or system (to continually read data).

`createSource` first [validates stream options](#).

`createSource` ...FIXME

## Validating General Options For Batch And Streaming Queries — `validateGeneralOptions` Internal Method

```
validateGeneralOptions(parameters: Map[String, String]): Unit
```

Note

Parameters are case-insensitive, i.e. `optionN` and `option` are equal.

`validateGeneralOptions` makes sure that exactly one topic subscription strategy is used in `parameters` and can be:

- `subscribe`
- `subscribepattern`
- `assign`

`validateGeneralOptions` reports an `IllegalArgumentException` when there is no subscription strategy in use or there are more than one strategies used.

`validateGeneralOptions` makes sure that the value of subscription strategies meet the requirements:

- `assign` strategy starts with `{` (the opening curly brace)
- `subscribe` strategy has at least one topic (in a comma-separated list of topics)
- `subscribepattern` strategy has the pattern defined

`validateGeneralOptions` makes sure that `group.id` has not been specified and reports an `IllegalArgumentException` otherwise.

Kafka option 'group.id' is not supported as user-specified consumer groups are not used to track offsets.

`validateGeneralOptions` makes sure that `auto.offset.reset` has not been specified and reports an `IllegalArgumentException` otherwise.

Kafka option 'auto.offset.reset' is not supported.  
Instead set the source option 'startingoffsets' to 'earliest' or 'latest' to specify where to start. Structured Streaming manages which offsets are consumed internally, rather than relying on the `kafkaConsumer` to do it. This will ensure that no data is missed when new topics/partitions are dynamically subscribed.  
Note that 'startingoffsets' only applies when a new Streaming query is started, and  
that resuming will always pick up from where the query left off.  
See the docs for more details.

`validateGeneralOptions` makes sure that the following options have not been specified and reports an `IllegalArgumentException` otherwise:

- `kafka.key.deserializer`
- `kafka.value.deserializer`
- `kafka.enable.auto.commit`
- `kafka.interceptor.classes`

In the end, `validateGeneralOptions` makes sure that `kafka.bootstrap.servers` option was specified and reports an `IllegalArgumentException` otherwise.

```
Option 'kafka.bootstrap.servers' must be specified for configuring Kafka consumer
```

**Note**

`validateGeneralOptions` is used when `KafkaSourceProvider` validates options for [streaming](#) and [batch](#) queries.

## Creating ConsumerStrategy — `strategy` Internal Method

```
strategy(caseInsensitiveParams: Map[String, String])
```

Internally, `strategy` finds the keys in the input `caseInsensitiveParams` that are one of the following and creates a corresponding [ConsumerStrategy](#).

Table 2. `KafkaSourceProvider.strategy`'s Key to ConsumerStrategy Conversion

Key	ConsumerStrategy
assign	<p><a href="#">AssignStrategy</a> with Kafka's <code>TopicPartitions</code>.</p> <hr/> <p><code>strategy</code> uses <code>JsonUtils.partitions</code> method to parse a JSON with topic names and partitions, e.g.</p> <pre>{"topicA": [0,1], "topicB": [0,1]}</pre> <p>The topic names and partitions are mapped directly to Kafka's <code>TopicPartition</code> objects.</p>
subscribe	<p><a href="#">SubscribeStrategy</a> with topic names</p> <hr/> <p><code>strategy</code> extracts topic names from a comma-separated string, e.g.</p> <pre>topic1,topic2,topic3</pre>
subscribepattern	<p><a href="#">SubscribePatternStrategy</a> with topic subscription regex pattern (that uses Java's <code>java.util.regex.Pattern</code> for the pattern), e.g.</p> <pre>topic\d</pre>

**Note**

`strategy` is used when:

- `KafkaSourceProvider` creates a `KafkaOffsetReader` for `KafkaSource`.
- `KafkaSourceProvider` creates a `KafkaRelation` (using `createRelation` method).

## Describing Streaming Source with Name and Schema

### — `sourceSchema` Method

```
sourceSchema(  
    sqlContext: SQLContext,  
    schema: Option[StructType],  
    providerName: String,  
    parameters: Map[String, String]): (String, StructType)
```

**Note**

`sourceSchema` is part of the [StreamSourceProvider Contract](#) to describe a [streaming source](#) with a name and the schema.

`sourceSchema` gives the [short name](#) (i.e. `kafka`) and the [fixed schema](#).

Internally, `sourceSchema` [validates Kafka options](#) and makes sure that the optional input `schema` is indeed undefined.

When the input `schema` is defined, `sourceSchema` reports a `IllegalArgumentException`.

Kafka source has a fixed schema and cannot be set with a custom one

## Validating Kafka Options for Streaming Queries

### — `validateStreamOptions` Internal Method

```
validateStreamOptions(caseInsensitiveParams: Map[String, String]): Unit
```

Firstly, `validateStreamOptions` makes sure that `endingoffsets` option is not used.

Otherwise, `validateStreamOptions` reports a `IllegalArgumentException`.

ending offset not valid in streaming queries

`validateStreamOptions` then [validates the general options](#).

**Note**

`validateStreamOptions` is used when `KafkaSourceProvider` is requested the [schema for Kafka source](#) and to [create a KafkaSource](#).

## Creating ContinuousReader for Continuous Stream Processing — `createContinuousReader` Method

```
createContinuousReader(  
    schema: Optional[StructType],  
    metadataPath: String,  
    options: DataSourceOptions): KafkaContinuousReader
```

**Note**

`createContinuousReader` is part of the [ContinuousReadSupport Contract](#) to create a [ContinuousReader](#).

`createContinuousReader` ...FIXME

## Converting Configuration Options to KafkaOffsetRangeLimit — `getKafkaOffsetRangeLimit` Object Method

```
getKafkaOffsetRangeLimit(  
    params: Map[String, String],  
    offsetOptionKey: String,  
    defaultOffsets: KafkaOffsetRangeLimit): KafkaOffsetRangeLimit
```

`getKafkaOffsetRangeLimit` finds the given `offsetOptionKey` in the `params` and does the following conversion:

- **latest** becomes [LatestOffsetRangeLimit](#)
- **earliest** becomes [EarliestOffsetRangeLimit](#)
- A JSON-formatted text becomes [SpecificOffsetRangeLimit](#)
- When the given `offsetOptionKey` is not found, `getKafkaOffsetRangeLimit` returns the given `defaultOffsets`

**Note**

`getKafkaOffsetRangeLimit` is used when `kafkaSourceProvider` is requested to [createSource](#), [createMicroBatchReader](#), [createContinuousReader](#), [createRelation](#), and [validateBatchOptions](#).

## Creating MicroBatchReader for Micro-Batch Stream Processing — `createMicroBatchReader` Method

```
createMicroBatchReader(  
    schema: Optional[StructType],  
    metadataPath: String,  
    options: DataSourceOptions): KafkaMicroBatchReader
```

**Note**

`createMicroBatchReader` is part of the [MicroBatchReadSupport Contract](#) to create a [MicroBatchReader](#) in [Micro-Batch Stream Processing](#).

`createMicroBatchReader` [validateStreamOptions](#) (in the given `DataSourceOptions`).

`createMicroBatchReader` generates a unique group ID of the format **spark-kafka-source-[randomUUID]-[metadataPath\_hashCode]** (to make sure that a new streaming query creates a new consumer group).

`createMicroBatchReader` finds all the parameters (in the given `DataSourceOptions`) that start with **kafka.** prefix, removes the prefix, and creates the current Kafka parameters.

`createMicroBatchReader` creates a [KafkaOffsetReader](#) with the following:

- [strategy](#) (in the given `DataSourceOptions`)
- [Properties for Kafka consumers on the driver](#) (given the current Kafka parameters, i.e. without **kafka.** prefix)
- The given `DataSourceOptions`
- **spark-kafka-source-[randomUUID]-[metadataPath\_hashCode]-driver** for the `driverGroupIdPrefix`

In the end, `createMicroBatchReader` creates a [KafkaMicroBatchReader](#) with the following:

- the `KafkaOffsetReader`
- [Properties for Kafka consumers on executors](#) (given the current Kafka parameters, i.e. without **kafka.** prefix) and the unique group ID (`spark-kafka-source-[randomUUID]-[metadataPath_hashCode]-driver`)
- The given `DataSourceOptions` and the `metadataPath`
- [Starting stream offsets](#) (`startingOffsets` option with the default of `LatestOffsetRangeLimit offsets`)
- [failOnDataLoss configuration property](#)

## Creating BaseRelation — `createRelation` Method

```
createRelation(  
    sqlContext: SQLContext,  
    parameters: Map[String, String]): BaseRelation
```

**Note**

`createRelation` is part of the [RelationProvider](#) contract to create a `BaseRelation`.

`createRelation` ...FIXME

## Validating Configuration Options for Batch Processing — `validateBatchOptions` Internal Method

```
validateBatchOptions(caseInsensitiveParams: Map[String, String]): Unit
```

`validateBatchOptions` ...FIXME

**Note**

`validateBatchOptions` is used exclusively when `KafkaSourceProvider` is requested to [createSource](#).

## kafkaParamsForDriver Method

```
kafkaParamsForDriver(specifiedKafkaParams: Map[String, String]): Map[String, Object]
```

`kafkaParamsForDriver` ...FIXME

**Note**

`kafkaParamsForDriver` is used when...FIXME

## kafkaParamsForExecutors Method

```
kafkaParamsForExecutors(  
    specifiedKafkaParams: Map[String, String],  
    uniqueGroupId: String): Map[String, Object]
```

`kafkaParamsForExecutors` sets the [Kafka properties for executors](#).

While setting the properties, `kafkaParamsForExecutors` prints out the following DEBUG message to the logs:

```
executor: Set [key] to [value], earlier value: [value]
```

**Note**

`kafkaParamsForExecutors` is used when:

- `KafkaSourceProvider` is requested to `createSource` (for a `KafkaSource`), `createMicroBatchReader` (for a `KafkaMicroBatchReader`), and `createContinuousReader` (for a `KafkaContinuousReader`)
- `KafkaRelation` is requested to `buildScan` (for a `KafkaSourceRDD`)

## Looking Up `failOnDataLoss` Configuration Property — `failOnDataLoss` Internal Method

```
failOnDataLoss(caseInsensitiveParams: Map[String, String]): Boolean
```

`failOnDataLoss` simply looks up the `failOnDataLoss` configuration property in the given `caseInsensitiveParams` (in case-insensitive manner) or defaults to `true`.

**Note**

`failOnDataLoss` is used when `KafkaSourceProvider` is requested to `createSource` (for a `KafkaSource`), `createMicroBatchReader` (for a `KafkaMicroBatchReader`), `createContinuousReader` (for a `KafkaContinuousReader`), and `createRelation` (for a `KafkaRelation`).

# KafkaSource

`KafkaSource` is a [streaming source](#) that generates `DataFrames` of records from one or more topics in Apache Kafka.

Note	Kafka topics are checked for new records every <a href="#">trigger</a> and so there is some noticeable delay between when the records have arrived to Kafka topics and when a Spark application processes them.
------	---

`KafkaSource` uses the [streaming metadata log directory](#) to persist offsets. The directory is the source ID under the `sources` directory in the `checkpointRoot` (of the `StreamExecution`).

Note	The <code>checkpointRoot</code> directory is one of the following: <ul style="list-style-type: none"> <li>• <code>checkpointLocation</code> option</li> <li>• <code>spark.sql.streaming.checkpointLocation</code> configuration property</li> </ul>
------	---

`KafkaSource` is created for [kafka](#) format (that is registered by [KafkaSourceProvider](#)).

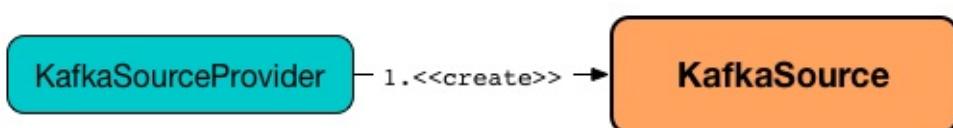


Figure 1. KafkaSource Is Created for kafka Format by KafkaSourceProvider

`KafkaSource` uses a [predefined \(fixed\) schema](#) (that cannot be changed).

`KafkaSource` also supports batch Datasets.

Tip	Enable <code>ALL</code> logging level for <code>org.apache.spark.sql.kafka010.KafkaSource</code> to see what happens inside. Add the following line to <code>conf/log4j.properties</code> : <pre>log4j.logger.org.apache.spark.sql.kafka010.KafkaSource=ALL</pre> Refer to <a href="#">Logging</a> .
-----	---

## Creating KafkaSource Instance

`KafkaSource` takes the following to be created:

- `SQLContext`
- `KafkaOffsetReader`

- Parameters of executors (reading from Kafka)
- Collection of key-value options
- **Streaming metadata log directory**, i.e. the directory for streaming metadata log (where `KafkaSource` persists `KafkaSourceOffset` offsets in JSON format)
- **Starting offsets** (as defined using `startingOffsets` option)
- Flag used to `create KafkaSourceRDDs` every trigger and when checking to `report a IllegalStateException` on data loss.

`KafkaSource` initializes the [internal properties](#).

## Generating Streaming DataFrame with Records From Kafka for Streaming Micro-Batch — `getBatch` Method

```
getBatch(
  start: Option[Offset],
  end: Offset): DataFrame
```

Note

`getBatch` is part of the [Source Contract](#) to generate a streaming `DataFrame` with data between the start and end `offsets`.

`getBatch` creates a streaming `DataFrame` with a query plan with `LogicalRDD` logical operator to scan data from a `KafkaSourceRDD`.

Internally, `getBatch` initializes [initial partition offsets](#) (unless initialized already).

You should see the following INFO message in the logs:

```
GetBatch called with start = [start], end = [end]
```

`getBatch` requests `KafkaSourceOffset` for `end partition offsets` for the input `end offset` (known as `untilPartitionOffsets` ).

`getBatch` requests `KafkaSourceOffset` for `start partition offsets` for the input `start offset` (if defined) or uses [initial partition offsets](#) (known as `fromPartitionOffsets` ).

`getBatch` finds the new partitions (as the difference between the topic partitions in `untilPartitionOffsets` and `fromPartitionOffsets` ) and requests `KafkaOffsetReader` to [fetch their earliest offsets](#).

`getBatch` [reports a data loss](#) if the new partitions don't match to what `KafkaOffsetReader` fetched.

```
Cannot find earliest offsets of [partitions]. Some data may have been missed
```

You should see the following INFO message in the logs:

```
Partitions added: [newPartitionOffsets]
```

`getBatch` [reports a data loss](#) if the new partitions don't have their offsets `0`.

```
Added partition [partition] starts from [offset] instead of 0. Some data may have been missed
```

`getBatch` [reports a data loss](#) if the `fromPartitionOffsets` partitions differ from `untilPartitionOffsets` partitions.

```
[partitions] are gone. Some data may have been missed
```

You should see the following DEBUG message in the logs:

```
TopicPartitions: [topicPartitions]
```

`getBatch` [gets the executors](#) (sorted by `executorId` and `host` of the registered block managers).

Important	That is when <code>getBatch</code> goes very low-level to allow for cached <code>KafkaConsumers</code> in the executors to be re-used to read the same partition in every batch (aka <i>location preference</i> ).
-----------	--

You should see the following DEBUG message in the logs:

```
Sorted executors: [sortedExecutors]
```

`getBatch` [creates a `KafkaSourceRDDOffsetRange`](#) per `TopicPartition`.

`getBatch` [filters out `KafkaSourceRDDOffsetRanges`](#) for which until offsets are smaller than from offsets. `getBatch` [reports a data loss](#) if they are found.

```
Partition [topicPartition]'s offset was changed from [fromOffset] to [untilOffset], so some data may have been missed
```

`getBatch` [creates a `KafkaSourceRDD`](#) (with `executorKafkaParams`, `pollTimeoutMs` and `reuseKafkaConsumer` flag enabled) and maps it to an RDD of `InternalRow`.

Important	<code>getBatch</code> creates a <code>KafkaSourceRDD</code> with <code>reuseKafkaConsumer</code> flag enabled.
-----------	--

You should see the following INFO message in the logs:

```
GetBatch generating RDD of offset range: [offsetRanges]
```

`getBatch` sets `currentPartitionOffsets` if it was empty (which is when...FIXME)

In the end, `getBatch` creates a streaming `DataFrame` for the `KafkaSourceRDD` and the schema.

## Fetching Offsets (From Metadata Log or Kafka Directly) — `getOffset` Method

```
getOffset: Option[Offset]
```

Note	<code>getOffset</code> is a part of the <a href="#">Source Contract</a> .
------	---

Internally, `getOffset` fetches the [initial partition offsets](#) (from the metadata log or Kafka directly).

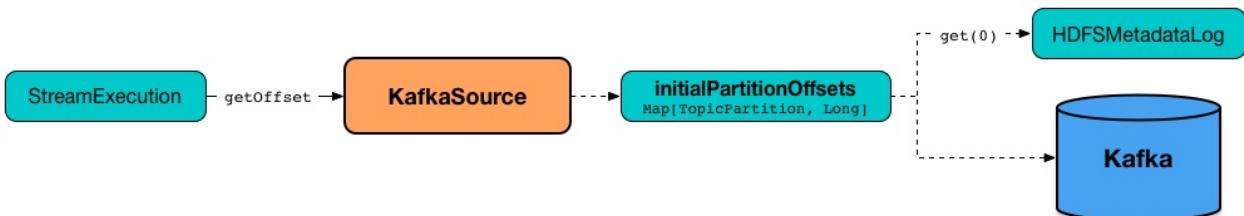


Figure 2. KafkaSource Initializing `initialPartitionOffsets` While Fetching Initial Offsets

Note	<code>initialPartitionOffsets</code> is a lazy value and is initialized the very first time <code>getOffset</code> is called (which is when <code>StreamExecution</code> constructs a streaming micro-batch).
------	---

```

scala> spark.version
res0: String = 2.3.0-SNAPSHOT

// Case 1: Checkpoint directory undefined
// initialPartitionOffsets read from Kafka directly
val records = spark.
  readStream.
  format("kafka").
  option("subscribe", "topic1").
  option("kafka.bootstrap.servers", "localhost:9092").
  load
// Start the streaming query
  
```

```

// dump records to the console every 10 seconds
import org.apache.spark.sql.streaming.{OutputMode, Trigger}
import scala.concurrent.duration._
val q = records.
  writeStream.
  format("console").
  option("truncate", false).
  trigger(Trigger.ProcessingTime(10.seconds)).
  outputMode(OutputMode.Update).
  start
// Note the temporary checkpoint directory
17/08/07 11:09:29 INFO StreamExecution: Starting [id = 75dd261d-6b62-40fc-a368-9d95d3c
b6f5f, runId = f18a5eb5-ccab-4d9d-8a81-befed41a72bd] with file:///private/var/folders/
0w/kb0d3rqn4zb9fcc91pxhgn8w0000gn/T/temporary-d0055630-24e4-4d9a-8f36-7a12a0f11bc0 to
store the query checkpoint.
...
INFO KafkaSource: Initial offsets: {"topic1":{"0":1}}

// Stop the streaming query
q.stop

// Case 2: Checkpoint directory defined
// initialPartitionOffsets read from Kafka directly
// since the checkpoint directory is not available yet
// it will be the next time the query is started
val records = spark.
  readStream.
  format("kafka").
  option("subscribe", "topic1").
  option("kafka.bootstrap.servers", "localhost:9092").
  load.
  select($"value" cast "string", $"topic", $"partition", $"offset")
import org.apache.spark.sql.streaming.{OutputMode, Trigger}
import scala.concurrent.duration._
val q = records.
  writeStream.
  format("console").
  option("truncate", false).
  option("checkpointLocation", "/tmp/checkpoint"). // <-- checkpoint directory
  trigger(Trigger.ProcessingTime(10.seconds)).
  outputMode(OutputMode.Update).
  start
// Note the checkpoint directory in use
17/08/07 11:21:25 INFO StreamExecution: Starting [id = b8f59854-61c1-4c2f-931d-62bbaf9
0ee3b, runId = 70d06a3b-f2b1-4fa8-a518-15df4cf59130] with file:///tmp/checkpoint to st
ore the query checkpoint.
...
INFO KafkaSource: Initial offsets: {"topic1":{"0":1}}
...
INFO StreamExecution: Stored offsets for batch 0. Metadata OffsetSeqMetadata(0,1502098
526848,Map(spark.sql.shuffle.partitions -> 200, spark.sql.streaming.stateStore.provide
rClass -> org.apache.spark.sql.execution.streaming.state.HDFSBackedStateStoreProvider)
)

```

```

// Review the checkpoint location
// $ ls -ltr /tmp/checkpoint/offsets
// total 8
// -rw-r--r-- 1 jacek wheel 248 7 sie 11:21 0
// $ tail -2 /tmp/checkpoint/offsets/0 | jq

// Produce messages to Kafka so the latest offset changes
// And more importantly the offset gets stored to checkpoint location
-----
Batch: 1
-----
+-----+-----+-----+
|value          |topic |partition|offset|
+-----+-----+-----+
|testing checkpoint location|topic1|0      |2      |
+-----+-----+-----+

// and one more
// Note the offset
-----
Batch: 2
-----
+-----+-----+-----+
|value      |topic |partition|offset|
+-----+-----+-----+
|another test|topic1|0      |3      |
+-----+-----+-----+

// See what was checkpointed
// $ ls -ltr /tmp/checkpoint/offsets
// total 24
// -rw-r--r-- 1 jacek wheel 248 7 sie 11:35 0
// -rw-r--r-- 1 jacek wheel 248 7 sie 11:37 1
// -rw-r--r-- 1 jacek wheel 248 7 sie 11:38 2
// $ tail -2 /tmp/checkpoint/offsets/2 | jq

// Stop the streaming query
q.stop

// And start over to see what offset the query starts from
// Checkpoint location should have the offsets
val q = records.
  writeStream.
  format("console").
  option("truncate", false).
  option("checkpointLocation", "/tmp/checkpoint"). // <-- checkpoint directory
  trigger(Trigger.ProcessingTime(10.seconds)).
  outputMode(OutputMode.Update).
  start
// Whoops...console format does not support recovery (!)
// Reported as https://issues.apache.org/jira/browse/SPARK-21667
org.apache.spark.sql.AnalysisException: This query does not support recovering from ch

```

```

eckpoint location. Delete /tmp/checkpoint/offsets to start over.;

  at org.apache.spark.sql.streaming.StreamingQueryManager.createQuery(StreamingQueryMa
nager.scala:222)
  at org.apache.spark.sql.streaming.StreamingQueryManager.startQuery(StreamingQueryMan
ager.scala:278)
  at org.apache.spark.sql.streaming.DataStreamWriter.start(DataStreamWriter.scala:284)
  ... 61 elided

// Change the sink (= output format) to JSON
val q = records.
  writeStream.
  format("json").
  option("path", "/tmp/json-sink").
  option("checkpointLocation", "/tmp/checkpoint"). // <-- checkpoint directory
  trigger(Trigger.ProcessingTime(10.seconds)).
  start
// Note the checkpoint directory in use
17/08/07 12:09:02 INFO StreamExecution: Starting [id = 02e00924-5f0d-4501-bcb8-80be8a8
be385, runId = 5eba2576-dad6-4f95-9031-e72514475edc] with file:///tmp/checkpoint to st
ore the query checkpoint.
...
17/08/07 12:09:02 INFO KafkaSource: GetBatch called with start = Some({topic1:{0:3
}}), end = {"topic1":{0:4}}
17/08/07 12:09:02 INFO KafkaSource: Partitions added: Map()
17/08/07 12:09:02 DEBUG KafkaSource: TopicPartitions: topic1-0
17/08/07 12:09:02 DEBUG KafkaSource: Sorted executors:
17/08/07 12:09:02 INFO KafkaSource: GetBatch generating RDD of offset range: KafkaSour
cerRDDOffsetRange(topic1-0, 3, 4, None)
17/08/07 12:09:03 DEBUG KafkaOffsetReader: Partitions assigned to consumer: [topic1-0]
. Seeking to the end.
17/08/07 12:09:03 DEBUG KafkaOffsetReader: Got latest offsets for partition : Map(topi
c1-0 -> 4)
17/08/07 12:09:03 DEBUG KafkaSource: GetOffset: ArrayBuffer((topic1-0,4))
17/08/07 12:09:03 DEBUG StreamExecution: getOffset took 122 ms
17/08/07 12:09:03 DEBUG StreamExecution: Resuming at batch 3 with committed offsets {K
afkaSource[Subscribe[topic1]]: {"topic1":{0:4}}} and available offsets {KafkaSource[Subsc
ribe[topic1]]: {"topic1":{0:4}}}
17/08/07 12:09:03 DEBUG StreamExecution: Stream running from {KafkaSource[Subscribe[t
opic1]]: {"topic1":{0:4}}} to {KafkaSource[Subscribe[topic1]]: {"topic1":{0:4}}}

```

`getOffset` requests `KafkaOffsetReader` to `fetchLatestOffsets` (known later as `latest`).

Note	(Possible performance degradation?) It is possible that <code>getOffset</code> will request the latest offsets from Kafka twice, i.e. while initializing <code>initialPartitionOffsets</code> (when no metadata log is available and <code>KafkaSource</code> 's <code>KafkaOffsetRangeLimit</code> is <code>LatestOffsetRangeLimit</code> ) and always as part of <code>getOffset</code> itself.
------	---

`getOffset` then calculates `currentPartitionOffsets` based on the `maxOffsetsPerTrigger` option.

Table 1. `getOffset's` Offset Calculation per `maxOffsetsPerTrigger`

<code>maxOffsetsPerTrigger</code>	<code>Offsets</code>
Unspecified (i.e. <code>None</code> )	<code>latest</code>
Defined (but <code>currentPartitionOffsets</code> is empty)	<code>rateLimit</code> with <code>limit</code> <code>limit</code> , <code>initialPartitionOffsets</code> as <code>from</code> , <code>until</code> as <code>latest</code>
Defined (and <code>currentPartitionOffsets</code> contains partitions and offsets)	<code>rateLimit</code> with <code>limit</code> <code>limit</code> , <code>currentPartitionOffsets</code> as <code>from</code> , <code>until</code> as <code>latest</code>

You should see the following DEBUG message in the logs:

```
DEBUG KafkaSource: GetOffset: [offsets]
```

In the end, `getOffset` creates a `KafkaSourceOffset` with `offsets` (as `Map[TopicPartition, Long]`).

## Fetching and Verifying Specific Offsets — `fetchAndVerify` Internal Method

```
fetchAndVerify(specificOffsets: Map[TopicPartition, Long]): KafkaSourceOffset
```

`fetchAndVerify` requests `KafkaOffsetReader` to `fetchSpecificOffsets` for the given `specificOffsets`.

`fetchAndVerify` makes sure that the starting offsets in `specificOffsets` are the same as in Kafka and [reports a data loss](#) otherwise.

```
startingOffsets for [tp] was [off] but consumer reset to [result(tp)]
```

In the end, `fetchAndVerify` creates a `KafkaSourceOffset` (with the result of [KafkaOffsetReader](#)).

Note	<code>fetchAndVerify</code> is used exclusively when <code>KafkaSource</code> initializes <a href="#">initial partition offsets</a> .
------	---

## Initial Partition Offsets (of 0th Batch) — `initialPartitionOffsets` Internal Lazy Property

```
initialPartitionOffsets: Map[TopicPartition, Long]
```

`initialPartitionOffsets` is the **initial partition offsets** for the batch `0` that were already persisted in the [streaming metadata log directory](#) or persisted on demand.

As the very first step, `initialPartitionOffsets` creates a custom [HDFSMetadataLog](#) (of `KafkaSourceOffsets` metadata) in the [streaming metadata log directory](#).

`initialPartitionOffsets` requests the `HDFSMetadataLog` for the [metadata](#) of the `0` th batch (as `KafkaSourceOffset` ).

If the metadata is available, `initialPartitionoffsets` requests the metadata for the collection of `TopicPartitions` and their offsets.

If the metadata could not be found, `initialPartitionOffsets` creates a new `KafkaSourceOffset` per [KafkaOffsetRangeLimit](#):

- For `EarliestOffsetRangeLimit`, `initialPartitionOffsets` requests the [KafkaOffsetReader](#) to [fetchEarliestOffsets](#)
- For `LatestOffsetRangeLimit`, `initialPartitionOffsets` requests the [KafkaOffsetReader](#) to [fetchLatestOffsets](#)
- For `SpecificOffsetRangeLimit`, `initialPartitionOffsets` requests the [KafkaOffsetReader](#) to [fetchSpecificOffsets](#) (and report a data loss per the `failOnDataLoss` flag)

`initialPartitionOffsets` requests the custom `HDFSMetadataLog` to [add the offsets to the metadata log](#) (as the metadata of the `0` th batch).

`initialPartitionOffsets` prints out the following INFO message to the logs:

```
Initial offsets: [offsets]
```

#### Note

`initialPartitionOffsets` is used when `KafkaSource` is requested for the following:

- Fetch offsets (from metadata log or Kafka directly)
- Generate a DataFrame with records from Kafka for a streaming batch (when the start offsets are not defined, i.e. before `StreamExecution` [commits the first streaming batch](#) and so nothing is in `committedOffsets` registry for a `KafkaSource` data source yet)

## HDFSMetadataLog.serialize

```
serialize(
  metadata: KafkaSourceOffset,
  out: OutputStream): Unit
```

Note	<code>serialize</code> is part of the <a href="#">HDFSMetadataLog Contract</a> to...FIXME.
------	--

`serialize` requests the `OutputStream` to write a zero byte (to support Spark 2.1.0 as per SPARK-19517).

`serialize` creates a `BufferedWriter` over a `OutputStreamWriter` over the `OutputStream` (with `UTF_8` charset encoding).

`serialize` requests the `BufferedWriter` to write the **v1** version indicator followed by a new line.

`serialize` then requests the `KafkaSourceOffset` for a JSON-serialized representation and the `BufferedWriter` to write it out.

In the end, `serialize` requests the `BufferedWriter` to flush (the underlying stream).

## rateLimit Internal Method

```
rateLimit(
  limit: Long,
  from: Map[TopicPartition, Long],
  until: Map[TopicPartition, Long]): Map[TopicPartition, Long]
```

`rateLimit` requests [KafkaOffsetReader](#) to [fetchEarliestOffsets](#).

Caution	FIXME
---------	-------

Note	<code>rateLimit</code> is used exclusively when <code>KafkaSource</code> gets available offsets (when <code>maxOffsetsPerTrigger</code> option is specified).
------	---

## getSortedExecutorList Method

Caution	FIXME
---------	-------

## reportDataLoss Internal Method

Caution	FIXME
---------	-------

Note	<p><code>reportDataLoss</code> is used when <code>kafkaSource</code> does the following:</p> <ul style="list-style-type: none"> <li>• fetches and verifies specific offsets</li> <li>• generates a DataFrame with records from Kafka for a batch</li> </ul>
------	---

## Internal Properties

Name	Description
<code>currentPartitionOffsets</code>	<p>Current partition offsets (as <code>Map[TopicPartition, Long]</code> )</p> <p>Initially <code>NONE</code> and set when <code>kafkaSource</code> is requested to get the maximum available offsets or generate a DataFrame with records from Kafka for a batch.</p>
<code>pollTimeoutMs</code>	
<code>sc</code>	<p>Spark Core's <code>SparkContext</code> (of the <code>SQLContext</code>)</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• Generating a DataFrame with records from Kafka for a streaming micro-batch (and creating a <code>KafkaSourceRDD</code>)</li> <li>• Initializing the <code>pollTimeoutMs</code> internal property</li> </ul>

# KafkaRelation

`KafkaRelation` represents a **collection of rows** with a [predefined schema](#) (`BaseRelation`) that supports [column pruning](#) (`TableScan`).

Tip	Read up on <a href="#">BaseRelation</a> and <a href="#">TableScan</a> in <a href="#">The Internals of Spark SQL</a> online book.
-----	--

`KafkaRelation` is [created](#) exclusively when `KafkaSourceProvider` is requested to [create a BaseRelation](#).

Table 1. KafkaRelation's Options

Name	Description
<code>kafkaConsumer.pollTimeoutMs</code>	Default: <code>spark.network.timeout</code> configuration if set or 120s
Tip	<p>Enable <code>ALL</code> logging levels for <code>org.apache.spark.sql.kafka010.KafkaRelation</code> to see what happens inside.</p> <p>Add the following line to <code>conf/log4j.properties</code> :</p> <pre style="background-color: #f0f0f0; padding: 5px;">log4j.logger.org.apache.spark.sql.kafka010.KafkaRelation=ALL</pre> <p>Refer to <a href="#">Logging</a>.</p>

## Creating KafkaRelation Instance

`KafkaRelation` takes the following when created:

- `SQLContext`
- [ConsumerStrategy](#)
- `Source options ( Map[String, String] )`
- User-defined Kafka parameters (`Map[String, String]`)
- `failOnDataLoss` flag
- [Starting offsets](#)
- [Ending offsets](#)

## getPartitionOffsets Internal Method

```
getPartitionOffsets(
    kafkaReader: KafkaOffsetReader,
    kafkaOffsets: KafkaOffsetRangeLimit): Map[TopicPartition, Long]
```

Caution

FIXME

Note

`getPartitionOffsets` is used exclusively when `KafkaRelation` builds RDD of rows (from the tuples).

## Building Distributed Data Scan with Column Pruning — buildScan Method

```
buildScan(): RDD[Row]
```

Note

`buildScan` is part of the `TableScan` contract to build a distributed data scan with column pruning.

`buildScan` generates a unique group ID of the format **spark-kafka-relation-[randomUUID]** (to make sure that a streaming query creates a new consumer group).

`buildScan` creates a `KafkaOffsetReader` with the following:

- The given `ConsumerStrategy` and the `source` options
- `Kafka parameters for the driver` based on the given `specifiedKafkaParams`
- **spark-kafka-relation-[randomUUID]-driver** for the `driverGroupIdPrefix`

`buildScan` uses the `KafkaOffsetReader` to `getPartitionOffsets` for the starting and ending offsets (based on the given `KafkaOffsetRangeLimit` and the `KafkaOffsetRangeLimit`, respectively). `buildScan` requests the `KafkaOffsetReader` to `close` afterwards.

`buildScan` creates offset ranges (that are a collection of `KafkaSourceRDDOffsetRanges` with a `Kafka TopicPartition`, beginning and ending offsets and undefined preferred location).

`buildScan` prints out the following INFO message to the logs:

```
Generating RDD of offset ranges: [offsetRanges]
```

`buildScan` creates a `KafkaSourceRDD` with the following:

- Kafka parameters for executors based on the given `specifiedKafkaParams` and the unique group ID ( `spark-kafka-relation-[randomUUID]` )
- The offset ranges created
- `pollTimeoutMs` configuration
- The given `failOnDataLoss` flag
- `reuseKafkaConsumer` flag off ( `false` )

`buildScan` requests the `KafkaSourceRDD` to map `Kafka ConsumerRecords` to `InternalRows`.

In the end, `buildScan` requests the `SQLContext` to create a `DataFrame` (with the name **kafka** and the predefined `schema`) that is immediately converted to a `RDD[InternalRow]`.

`buildScan` throws a `IllegalStateException` when...FIXME

```
different topic partitions for starting offsets topics[[fromTopics]] and ending offset  
s topics[[untilTopics]]
```

`buildScan` throws a `IllegalStateException` when...FIXME

```
[tp] doesn't have a from offset
```

# KafkaSourceRDD

`KafkaSourceRDD` is an `RDD` of Kafka's `ConsumerRecords` (`RDD[ConsumerRecord[Array[Byte], Array[Byte]]]`) and no parent RDDs.

`KafkaSourceRDD` is created when:

- `KafkaRelation` is requested to build a distributed data scan with column pruning
- `KafkaSource` is requested to generate a streaming DataFrame with records from Kafka for a streaming micro-batch

## Creating KafkaSourceRDD Instance

`KafkaSourceRDD` takes the following when created:

- `SparkContext`
- Collection of key-value settings for executors reading records from Kafka topics
- Collection of `KafkaSourceRDDOffsetRange` offsets
- Timeout (in milliseconds) to poll data from Kafka

Used when `KafkaSourceRDD` is requested for records (for given offsets) and in turn requests `CachedKafkaConsumer` to poll for Kafka's `ConsumerRecords`.

- Flag to...FIXME
- Flag to...FIXME

## Placement Preferences of Partition (Preferred Locations)

### — `getPreferredLocations` Method

```
getPreferredLocations(  
    split: Partition): Seq[String]
```

Note	<code>getPreferredLocations</code> is part of the <code>RDD</code> contract to specify placement preferences.
------	---

`getPreferredLocations` converts the given `Partition` to a `KafkaSourceRDDPartition` and...  
FIXME

## Computing Partition— `compute` Method

```
compute(
  thePart: Partition,
  context: TaskContext
): Iterator[ConsumerRecord[Array[Byte], Array[Byte]]]
```

Note

`compute` is part of the `RDD` contract to compute a given partition.

`compute` uses `KafkaDataConsumer` utility to [acquire a cached KafkaDataConsumer](#) (for a partition).

`compute` [resolves the range](#) (based on the `offsetRange`) of the given partition that is assumed a `KafkaSourceRDDPartition`.

`compute` returns a `NextIterator` so that `getNext` uses the `KafkaDataConsumer` to [get a record](#).

When the beginning and ending offsets (of the offset range) are equal, `compute` prints out the following INFO message to the logs, requests the `KafkaDataConsumer` to [release](#) and returns an empty iterator.

```
Beginning offset [fromOffset] is the same as ending offset skipping [topic] [partition]
```

`compute` throws an `AssertionError` when the beginning offset (`fromOffset`) is after the ending offset (`untilOffset`):

```
Beginning offset [fromOffset] is after the ending offset
[untilOffset] for topic [topic] partition [partition]. You
either provided an invalid fromOffset, or the Kafka topic has
been damaged
```

## getPartitions Method

```
getPartitions: Array[Partition]
```

Note

`getPartitions` is part of the `RDD` contract to...FIXME.

`getPartitions` ...FIXME

## Persisting RDD — `persist` Method

```
persist: Array[Partition]
```

Note

`persist` is part of the `RDD` contract to persist an RDD.

`persist ...FIXME`

## `resolveRange` Internal Method

```
resolveRange(  
    consumer: KafkaDataConsumer,  
    range: KafkaSourceRDDOffsetRange  
) : KafkaSourceRDDOffsetRange
```

`resolveRange ...FIXME`

Note

`resolveRange` is used when...FIXME

# CachedKafkaConsumer

Caution	FIXME
---------	-------

## **poll Internal Method**

Caution	FIXME
---------	-------

## **fetchData Internal Method**

Caution	FIXME
---------	-------

# KafkaSourceOffset

`KafkaSourceOffset` is a custom [Offset](#) for [kafka data source](#).

`KafkaSourceOffset` is [created](#) (directly or indirectly using [apply](#)) when:

- `KafkaContinuousReader` is requested to [setStartOffset](#), [deserializeOffset](#), and [mergeOffsets](#)
- `KafkaMicroBatchReader` is requested to [getStartOffset](#), [getEndOffset](#), [deserializeOffset](#), and [getOrCreateInitialPartitionOffsets](#)
- `KafkaOffsetReader` is requested to [fetchSpecificOffsets](#)
- `KafkaSource` is requested for the [initial partition offsets](#) (of 0th batch) and [getOffset](#)
- `KafkaSourceInitialOffsetWriter` is requested to [deserialize](#) a `KafkaSourceOffset` (from an [InputStream](#))
- `KafkaSourceOffset` is requested for [partition offsets](#)

`KafkaSourceOffset` takes a collection of Kafka `TopicPartitions` with offsets to be created.

## Partition Offsets — `getPartitionOffsets` Method

```
getPartitionOffsets(  
    offset: Offset): Map[TopicPartition, Long]
```

`getPartitionOffsets` takes [KafkaSourceOffset.partitionToOffsets](#) from `offset`.

If `offset` is `KafkaSourceOffset`, `getPartitionOffsets` takes the partitions and offsets straight from it.

If however `offset` is `SerializedOffset`, `getPartitionOffsets` deserializes the offsets from JSON.

`getPartitionOffsets` reports an `IllegalArgumentException` when `offset` is neither `KafkaSourceOffset` or `SerializedOffset`.

```
Invalid conversion from offset of [class] to KafkaSourceOffset
```

**Note**

- `getPartitionOffsets` is used when:
- `KafkaContinuousReader` is requested to [planInputPartitions](#)
  - `KafkaSource` is requested to [generate a streaming DataFrame with records from Kafka for a streaming micro-batch](#)

**JSON-Encoded Offset — `json` Method**

```
json: String
```

**Note**

`json` is part of the [Offset Contract](#) for a JSON-encoded offset.

```
json ...FIXME
```

**Creating KafkaSourceOffset Instance — `apply` Utility Method**

```
apply(  
    offsetTuples: (String, Int, Long)*): KafkaSourceOffset (1)  
apply(  
    offset: SerializedOffset): KafkaSourceOffset
```

1. Used in tests only

```
apply ...FIXME
```

**Note**

- `apply` is used when:
- `KafkaSourceInitialOffsetWriter` is requested to [deserialize a KafkaSourceOffset \(from an InputStream\)](#)
  - `KafkaSource` is requested for the [initial partition offsets \(of 0th batch\)](#)
  - `KafkaSourceOffset` is requested to [getPartitionOffsets](#)

# KafkaOffsetReader

`KafkaOffsetReader` relies on the [ConsumerStrategy](#) to create a Kafka Consumer.

`KafkaOffsetReader` creates a Kafka Consumer with `group.id`

(`ConsumerConfig.GROUP_ID_CONFIG`) configuration explicitly set to `nextGroupId` (i.e. the given `driverGroupIdPrefix` followed by `nextId`).

`KafkaOffsetReader` is created when:

- `KafkaRelation` is requested to build a distributed data scan with column pruning
- `KafkaSourceProvider` is requested to create a `KafkaSource`, `createMicroBatchReader`, and `createContinuousReader`

Table 1. KafkaOffsetReader's Options

Name	Description
<code>fetchOffset.numRetries</code>	Default: 3
<code>fetchOffset.retryIntervalMs</code>	How long to wait before retries Default: 1000

`KafkaOffsetReader` defines the [predefined fixed schema](#).

<b>Tip</b>	<p>Enable <code>ALL</code> logging level for <code>org.apache.spark.sql.kafka010.KafkaOffsetReader</code> to see what happens inside.</p> <p>Add the following line to <code>conf/log4j.properties</code> :</p> <pre>log4j.logger.org.apache.spark.sql.kafka010.KafkaOffsetReader=ALL</pre> <p>Refer to <a href="#">Logging</a>.</p>
------------	--

## Creating KafkaOffsetReader Instance

`KafkaOffsetReader` takes the following to be created:

- [ConsumerStrategy](#)
- Kafka parameters (as name-value pairs that are used exclusively to [create a Kafka consumer](#))
- Options (as name-value pairs)

- Prefix of the group ID

`KafkaOffsetReader` initializes the [internal properties](#).

## nextGroupId Internal Method

```
nextGroupId(): String
```

`nextGroupId` sets the [groupId](#) to be the [driverGroupIdPrefix](#), - followed by the [nextId](#) (i.e. `[driverGroupIdPrefix]-[nextId]` ).

In the end, `nextGroupId` increments the [nextId](#) and returns the [groupId](#).

Note

`nextGroupId` is used exclusively when `KafkaOffsetReader` is requested for a [Kafka Consumer](#).

## resetConsumer Internal Method

```
resetConsumer(): Unit
```

`resetConsumer` ...FIXME

Note

`resetConsumer` is used when...FIXME

## fetchTopicPartitions Method

```
fetchTopicPartitions(): Set[TopicPartition]
```

Caution

FIXME

Note

`fetchTopicPartitions` is used when `KafkaRelation` [getPartitionOffsets](#).

## Fetching Earliest Offsets — `fetchEarliestOffsets` Method

```
fetchEarliestOffsets(): Map[TopicPartition, Long]
fetchEarliestOffsets(newPartitions: Seq[TopicPartition]): Map[TopicPartition, Long]
```

Caution

FIXME

## Note

`fetchEarliestOffsets` is used when `KafkaSource` `rateLimit` and `generates a DataFrame for a batch` (when new partitions have been assigned).

## Fetching Latest Offsets — `fetchLatestOffsets` Method

```
fetchLatestOffsets(): Map[TopicPartition, Long]
```

## Caution

FIXME

## Note

`fetchLatestOffsets` is used when `KafkaSource` `gets offsets` or `initialPartitionOffsets` is `initialized`.

## `withRetriesWithoutInterrupt` Internal Method

```
withRetriesWithoutInterrupt(  
  body: => Map[TopicPartition, Long]): Map[TopicPartition, Long]
```

```
withRetriesWithoutInterrupt ...FIXME
```

## Note

`withRetriesWithoutInterrupt` is used when...FIXME

## Fetching Offsets for Selected TopicPartitions — `fetchSpecificOffsets` Method

```
fetchSpecificOffsets(  
  partitionOffsets: Map[TopicPartition, Long],  
  reportDataLoss: String => Unit): KafkaSourceOffset
```

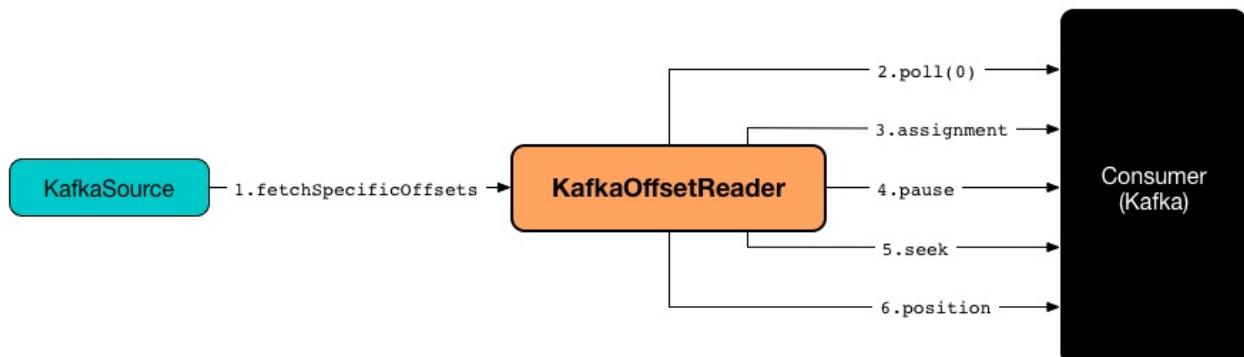


Figure 1. KafkaOffsetReader's `fetchSpecificOffsets`

`fetchSpecificOffsets` requests the `Kafka Consumer` to `poll(0)`.

`fetchSpecificOffsets` requests the Kafka Consumer for assigned partitions (using `Consumer.assignment()`).

`fetchSpecificOffsets` requests the Kafka Consumer to `pause(partitions)`.

You should see the following DEBUG message in the logs:

```
DEBUG KafkaOffsetReader: Partitions assigned to consumer: [partitions]. Seeking to [partitionOffsets]
```

For every partition offset in the input `partitionOffsets`, `fetchSpecificOffsets` requests the Kafka Consumer to:

- `seekToEnd` for the latest (aka `-1`)
- `seekToBeginning` for the earliest (aka `-2`)
- `seek` for other offsets

In the end, `fetchSpecificOffsets` creates a collection of Kafka's `TopicPartition` and `position` (using the Kafka Consumer).

Note	<code>fetchSpecificOffsets</code> is used when <code>KafkaSource</code> fetches and verifies initial partition offsets.
------	---

## Creating Kafka Consumer — `createConsumer` Internal Method

```
createConsumer(): Consumer[Array[Byte], Array[Byte]]
```

`createConsumer` requests `ConsumerStrategy` to create a Kafka Consumer with `driverKafkaParams` and new generated group.id Kafka property.

Note	<code>createConsumer</code> is used when <code>KafkaOffsetReader</code> is created (and initializes consumer) and <code>resetConsumer</code>
------	--

## Creating Kafka Consumer (Unless Already Available) — `consumer` Method

```
consumer: Consumer[Array[Byte], Array[Byte]]
```

`consumer` gives the cached Kafka Consumer or creates one itself.

## Note

Since `consumer` method is used (to access the internal Kafka Consumer) in the `fetch` methods that gives the property of creating a new Kafka Consumer whenever the internal Kafka Consumer reference becomes `null`, i.e. as in `resetConsumer`.

`consumer ...FIXME`

## Note

`consumer` is used when `KafkaOffsetReader` is requested to `fetchTopicPartitions`, `fetchSpecificOffsets`, `fetchEarliestOffsets`, and `fetchLatestOffsets`.

**Closing — close Method**

`close(): Unit`

`close` stop the Kafka Consumer (if the Kafka Consumer is available).

`close` requests the ExecutorService to shut down.

## Note

`close` is used when:

- `KafkaContinuousReader`, `KafkaMicroBatchReader`, and `KafkaSource` are requested to stop a streaming reader or source
- `KafkaRelation` is requested to build a distributed data scan with column pruning

**runUninterruptibly Internal Method**

`runUninterruptibly[T](body: => T): T`

`runUninterruptibly ...FIXME`

## Note

`runUninterruptibly` is used when...FIXME

**stopConsumer Internal Method**

`stopConsumer(): Unit`

`stopConsumer ...FIXME`

## Note

`stopConsumer` is used when...FIXME

## Textual Representation — `toString` Method

`toString: String`

Note

`toString` is part of the [java.lang.Object](#) contract for the string representation of the object.

`toString` ...FIXME

## Internal Properties

Name	Description
<code>_consumer</code>	Kafka's <a href="#">Consumer</a> ( <code>Consumer[Array[Byte], Array[Byte]]</code> ) Initialized when <code>KafkaOffsetReader</code> is created. Used when <code>KafkaOffsetReader</code> : <ul style="list-style-type: none"><li>• <code>fetchTopicPartitions</code></li><li>• fetches offsets for selected TopicPartitions</li><li>• <code>fetchEarliestOffsets</code></li><li>• <code>fetchLatestOffsets</code></li><li>• <code>resetConsumer</code></li><li>• <code>is closed</code></li></ul>
<code>execContext</code>	<a href="#">scala.concurrent.ExecutionContextExecutorService</a>
<code>groupId</code>	
<code>kafkaReaderThread</code>	<a href="#">java.util.concurrent.ExecutorService</a>
<code>maxOffsetFetchAttempts</code>	
<code>nextId</code>	Initially <code>0</code>
<code>offsetFetchAttemptIntervalMs</code>	

# ConsumerStrategy Contract for KafkaConsumer Providers

`ConsumerStrategy` is the [contract](#) for components that can [create a KafkaConsumer](#) using the given Kafka parameters.

```
createConsumer(kafkaParams: java.util.Map[String, Object]): Consumer[Array[Byte], Array[Byte]]
```

Table 1. Available ConsumerStrategies

ConsumerStrategy	createConsumer		
AssignStrategy	Uses <a href="#">KafkaConsumer.assign(Collection&lt;TopicPartition&gt; partitions)</a>		
SubscribeStrategy	Uses <a href="#">KafkaConsumer.subscribe(Collection&lt;String&gt; topics)</a>		
SubscribePatternStrategy	Uses <a href="#">KafkaConsumer.subscribe(Pattern pattern, ConsumerRebalanceListener listener)</a> with <code>NoOpConsumerRebalanceListener</code> . <table border="1" style="margin-left: 20px;"> <tr> <td style="padding: 5px;">Tip</td> <td style="padding: 5px;">Refer to <a href="#">java.util.regex.Pattern</a> for the format of supported topic subscription regex patterns.</td> </tr> </table>	Tip	Refer to <a href="#">java.util.regex.Pattern</a> for the format of supported topic subscription regex patterns.
Tip	Refer to <a href="#">java.util.regex.Pattern</a> for the format of supported topic subscription regex patterns.		

# KafkaSink

`KafkaSink` is a [streaming sink](#) that [KafkaSourceProvider](#) registers as the `kafka` format.

```
// start spark-shell or a Spark application with spark-sql-kafka-0-10 module
// spark-shell --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.3.0-SNAPSHOT
import org.apache.spark.sql.SparkSession
val spark: SparkSession = ...
spark.
  readStream.
  format("text").
  load("server-logs/*.out").
  as[String].
  writeStream.
  queryName("server-logs processor").
  format("kafka"). // <-- uses KafkaSink
  option("topic", "topic1").
  option("checkpointLocation", "/tmp/kafka-sink-checkpoint"). // <-- mandatory
  start

// in another terminal
$ echo hello > server-logs/hello.out

// in the terminal with Spark
FIXME
```

## Creating KafkaSink Instance

`KafkaSink` takes the following when created:

- `SQLContext`
- Kafka parameters (used on executor) as a map of `(String, Object)` pairs
- Optional topic name

## addBatch Method

```
addBatch(batchId: Long, data: DataFrame): Unit
```

Internally, `addBatch` requests `KafkaWriter` to write the input `data` to the [topic](#) (if defined) or a topic in [executorKafkaParams](#).

Note	<code>addBatch</code> is a part of <a href="#">Sink Contract</a> to "add" a batch of data to the sink.
------	--



# KafkaOffsetRangeLimit — Desired Offset Range Limits

`KafkaOffsetRangeLimit` represents the desired offset range limits for starting, ending, and specific offsets in [Kafka Data Source](#).

Table 1. KafkaOffsetRangeLimits

KafkaOffsetRangeLimit	Description
<code>EarliestOffsetRangeLimit</code>	Intent to bind to the <b>earliest</b> offset
<code>LatestOffsetRangeLimit</code>	Intent to bind to the <b>latest</b> offset
<code>SpecificOffsetRangeLimit</code>	Intent to bind to <b>specific offsets</b> with the following special offset "magic" numbers: <ul style="list-style-type: none"> <li>• <code>-1</code> or <code>KafkaOffsetRangeLimit.LATEST</code> - the latest offset</li> <li>• <code>-2</code> or <code>KafkaOffsetRangeLimit.EARLIEST</code> - the earliest offset</li> </ul>

Note

`KafkaOffsetRangeLimit` is a Scala **sealed trait** which means that all the [implementations](#) are in the same compilation unit (a single file).

`KafkaOffsetRangeLimit` is often used in a text-based representation and is converted to from **latest**, **earliest** or a **JSON-formatted text** using [KafkaSourceProvider.getKafkaOffsetRangeLimit](#) object method.

Note

A JSON-formatted text is of the following format `{"topicName": {"partition":offset},...}` , e.g. `{"topicA":{"0":23,"1":-1}, "topicB":{"0":-2}}` .

`KafkaOffsetRangeLimit` is used when:

- [KafkaContinuousReader](#) is created (with the [initial offsets](#))
- [KafkaMicroBatchReader](#) is created (with the [starting offsets](#))
- [KafkaRelation](#) is created (with the [starting](#) and [ending](#) offsets)
- [KafkaSource](#) is created (with the [starting offsets](#))
- `KafkaSourceProvider` is requested to [convert configuration options](#) to [KafkaOffsetRangeLimits](#)



# KafkaDataConsumer

`KafkaDataConsumer` is the abstraction of [Kafka consumers](#) that use [InternalKafkaConsumer](#) that can be [released](#).

Table 1. KafkaDataConsumer Contract (Abstract Methods Only)

Method	Description
<code>internalConsumer</code>	<code>internalConsumer: InternalKafkaConsumer</code> Used when...FIXME
<code>release</code>	<code>release(): Unit</code> Used when...FIXME

Table 2. KafkaDataConsumers

KafkaDataConsumer	Description
<code>CachedKafkaDataConsumer</code>	
<code>NonCachedKafkaDataConsumer</code>	

## Acquiring Cached KafkaDataConsumer for Partition — `acquire` Object Method

```
acquire(
    topicPartition: TopicPartition,
    kafkaParams: ju.Map[String, Object],
    useCache: Boolean
): KafkaDataConsumer
```

`acquire ...FIXME`

Note

`acquire` is used when...FIXME

## Getting Kafka Record — `get` Method

```
get(  
    offset: Long,  
    untilOffset: Long,  
    pollTimeoutMs: Long,  
    failOnDataLoss: Boolean  
) : ConsumerRecord[Array[Byte], Array[Byte]]
```

get ...FIXME

Note

get is used when...FIXME

# KafkaMicroBatchReader

`KafkaMicroBatchReader` is the [MicroBatchReader](#) for `kafka` data source for [Micro-Batch Stream Processing](#).

`KafkaMicroBatchReader` is [created](#) exclusively when `KafkaSourceProvider` is requested to [create a MicroBatchReader](#).

`KafkaMicroBatchReader` uses the [DataSourceOptions](#) to access the `kafkaConsumer.pollTimeoutMs` option (default: `spark.network.timeout` or `120s`).

`KafkaMicroBatchReader` uses the [DataSourceOptions](#) to access the `maxOffsetsPerTrigger` option (default: `(undefined)`).

`KafkaMicroBatchReader` uses the [Kafka properties for executors](#) to create [KafkaMicroBatchInputPartitions](#) when requested to `planInputPartitions`.

<b>Tip</b>	<p>Enable <code>ALL</code> logging level for <code>org.apache.spark.sql.kafka010.KafkaMicroBatchReader</code> to see what happens inside.</p> <p>Add the following line to <code>conf/log4j.properties</code> :</p> <pre>log4j.logger.org.apache.spark.sql.kafka010.KafkaMicroBatchReader=ALL</pre> <p>Refer to <a href="#">Logging</a>.</p>
------------	--

## Creating KafkaMicroBatchReader Instance

`KafkaMicroBatchReader` takes the following to be created:

- [KafkaOffsetReader](#)
- Kafka properties for executors (`Map[String, Object]`)
- `DataSourceOptions`
- Metadata Path
- Desired starting [KafkaOffsetRangeLimit](#)
- `failOnDataLoss` option

`KafkaMicroBatchReader` initializes the internal registries and counters.

## readSchema Method

```
readSchema(): StructType
```

**Note** `readSchema` is part of the `DataSourceReader` contract to...FIXME.

`readSchema` simply returns the [predefined fixed schema](#).

## Stopping Streaming Reader — stop Method

```
stop(): Unit
```

**Note** `stop` is part of the [BaseStreamingSource Contract](#) to stop a streaming reader.

`stop` simply requests the [KafkaOffsetReader](#) to [close](#).

## Plan Input Partitions — planInputPartitions Method

```
planInputPartitions(): java.util.List[InputPartition[InternalRow]]
```

**Note** `planInputPartitions` is part of the `DataSourceReader` contract in Spark SQL for the number of `InputPartitions` to use as RDD partitions (when `DataSourceV2ScanExec` physical operator is requested for the partitions of the input RDD).

`planInputPartitions` first finds the new partitions (`TopicPartitions` that are in the `endPartitionOffsets` but not in the `startPartitionOffsets`) and requests the [KafkaOffsetReader](#) to [fetch their earliest offsets](#).

`planInputPartitions` prints out the following INFO message to the logs:

```
Partitions added: [newPartitionInitialOffsets]
```

`planInputPartitions` then prints out the following DEBUG message to the logs:

```
TopicPartitions: [comma-separated list of TopicPartitions]
```

`planInputPartitions` requests the [KafkaOffsetRangeCalculator](#) for [offset ranges](#) (given the `startPartitionOffsets` and the newly-calculated `newPartitionInitialOffsets` as the `fromOffsets`, the `endPartitionOffsets` as the `untilOffsets`, and the [available executors](#)

(sorted in descending order)).

In the end, `planInputPartitions` creates a [KafkaMicroBatchInputPartition](#) for every offset range (with the [Kafka properties for executors](#), the `pollTimeoutMs`, the `failOnDataLoss` flag and whether to reuse a Kafka consumer among Spark tasks).

Note	<a href="#">KafkaMicroBatchInputPartition</a> uses a shared Kafka consumer only when all the offset ranges have distinct <code>TopicPartitions</code> , so concurrent tasks (of a stage in a Spark job) will not interfere and read the same <code>TopicPartitions</code> .
------	---

`planInputPartitions` [reports data loss](#) when...FIXME

## Available Executors in Spark Cluster (Sorted By Host and Executor ID in Descending Order)

### — `getSortedExecutorList` Internal Method

	<code>getSortedExecutorList(): Array[String]</code>
--	---

`getSortedExecutorList` requests the `BlockManager` to request the `BlockManagerMaster` to get the peers (the other nodes in a Spark cluster), creates a `ExecutorCacheTaskLocation` for every pair of host and executor ID, and in the end, sort it in descending order.

Note	<code>getSortedExecutorList</code> is used exclusively when <code>KafkaMicroBatchReader</code> is requested to <a href="#">planInputPartitions</a> (and calculates offset ranges).
------	--

## `getOrCreateInitialPartitionOffsets` Internal Method

	<code>getOrCreateInitialPartitionOffsets(): PartitionOffsetMap</code>
--	---

`getOrCreateInitialPartitionOffsets` ...FIXME

Note	<code>getOrCreateInitialPartitionOffsets</code> is used exclusively for the <a href="#">initialPartitionOffsets</a> internal registry.
------	--

## `getStartOffset` Method

	<code>getStartOffset: Offset</code>
--	-------------------------------------

Note	<code>getStartOffset</code> is part of the <a href="#">MicroBatchReader Contract</a> to get the start (beginning) offsets.
------	--

```
getStartOffset ...FIXME
```

## getEndOffset Method

```
getEndOffset: Offset
```

**Note** `getEndOffset` is part of the [MicroBatchReader Contract](#) to get the end `offsets`.

```
getEndOffset ...FIXME
```

## deserializeOffset Method

```
deserializeOffset(json: String): Offset
```

**Note** `deserializeOffset` is part of the [MicroBatchReader Contract](#) to deserialize an `offset` (from JSON format).

```
deserializeOffset ...FIXME
```

## Internal Properties

Name	Description
<code>endPartitionOffsets</code>	Ending offsets for the assigned partitions ( <code>Map[TopicPartition, Long]</code> )  Used when...FIXME
<code>initialPartitionOffsets</code>	<code>initialPartitionOffsets: Map[TopicPartition, Long]</code>
<code>rangeCalculator</code>	<a href="#">KafkaOffsetRangeCalculator</a> (for the given <a href="#">DataSourceOptions</a> )  Used exclusively when <code>KafkaMicroBatchReader</code> is requested to <a href="#">planInputPartitions</a> (to calculate offset ranges)
<code>startPartitionOffsets</code>	Starting offsets for the assigned partitions ( <code>Map[TopicPartition, Long]</code> )  Used when...FIXME



# KafkaOffsetRangeCalculator

`KafkaOffsetRangeCalculator` is created for `KafkaMicroBatchReader` to calculate offset ranges (when `KafkaMicroBatchReader` is requested to `planInputPartitions`).

`KafkaOffsetRangeCalculator` takes an optional **minimum number of partitions per executor** (`minPartitions`) to be created (that can either be undefined or greater than `0`).

When created with a `DataSourceOptions`, `KafkaOffsetRangeCalculator` uses `minPartitions` option for the **minimum number of partitions per executor**.

## Offset Ranges — `getRanges` Method

```
getRanges(
    fromOffsets: PartitionOffsetMap,
    untilOffsets: PartitionOffsetMap,
    executorLocations: Seq[String] = Seq.empty): Seq[KafkaOffsetRange]
```

`getRanges` finds the common `TopicPartitions` that are the keys that are used in the `fromOffsets` and `untilOffsets` collections (*intersection*).

For every common `TopicPartition`, `getRanges` creates a `KafkaOffsetRange` with the from and until offsets from the `fromOffsets` and `untilOffsets` collections (and the `preferredLoc` undefined). `getRanges` filters out the `TopicPartitions` that have no records to consume (i.e. the difference between until and from offsets is not greater than `0`).

At this point, `getRanges` knows the `TopicPartitions` with records to consume.

`getRanges` branches off based on the defined **minimum number of partitions per executor** and the number of `KafkaOffsetRanges` (`TopicPartitions` with records to consume).

For the **minimum number of partitions per executor** undefined or smaller than the number of `KafkaOffsetRanges` (`TopicPartitions` to consume records from), `getRanges` updates every `KafkaOffsetRange` with the `preferred executor` based on the `TopicPartition` and the `executorLocations` ).

Otherwise (with the **minimum number of partitions per executor** defined and greater than the number of `KafkaOffsetRanges` ), `getRanges` splits `KafkaOffsetRanges` into smaller ones.

Note	<code>getRanges</code> is used exclusively when <code>KafkaMicroBatchReader</code> is requested to <code>planInputPartitions</code> .
------	---

## KafkaOffsetRange — TopicPartition with From and Until Offsets and Optional Preferred Location

`KafkaOffsetRange` is a case class with the following attributes:

- `TopicPartition`
- `fromOffset offset`
- `untilOffset offset`
- Optional preferred location

`KafkaOffsetRange` knows the size, i.e. the number of records between the `untilOffset` and `fromOffset` offsets.

## Selecting Preferred Executor for TopicPartition — `getLocation` Internal Method

```
getLocation(  
    tp: TopicPartition,  
    executorLocations: Seq[String]): Option[String]
```

`getLocation` ...FIXME

Note

`getLocation` is used exclusively when `KafkaOffsetRangeCalculator` is requested to calculate offset ranges.

# KafkaMicroBatchInputPartition

`KafkaMicroBatchInputPartition` is an `InputPartition` (of `InternalRows`) that is used (created) exclusively when `KafkaMicroBatchReader` is requested for `input partitions` (when `DataSourceV2ScanExec` physical operator is requested for the partitions of the input RDD).

`KafkaMicroBatchInputPartition` takes the following to be created:

- `KafkaOffsetRange`
- Kafka parameters used for Kafka clients on executors (`Map[String, Object]`)
- Poll timeout (in ms)
- `failOnDataLoss` flag
- `reuseKafkaConsumer` flag

`KafkaMicroBatchInputPartition` creates a `KafkaMicroBatchInputPartitionReader` when requested for a `InputPartitionReader[InternalRow]` (as a part of the `InputPartition` contract).

`KafkaMicroBatchInputPartition` simply requests the given `KafkaOffsetRange` for the optional `preferredLoc` when requested for `preferredLocations` (as a part of the `InputPartition` contract).

# KafkaMicroBatchInputPartitionReader

`KafkaMicroBatchInputPartitionReader` is an `InputPartitionReader` (of `InternalRows`) that is created exclusively when `KafkaMicroBatchInputPartition` is requested for `one` (as a part of the `InputPartition` contract).

## Creating KafkaMicroBatchInputPartitionReader Instance

`KafkaMicroBatchInputPartitionReader` takes the following to be created:

- `KafkaOffsetRange`
- Kafka parameters used for Kafka clients on executors (`Map[String, Object]`)
- Poll timeout (in ms)
- `failOnDataLoss` flag
- `reuseKafkaConsumer` flag

Note	All the input arguments to create a <code>KafkaMicroBatchInputPartitionReader</code> are exactly the input arguments used to create a <code>KafkaMicroBatchInputPartition</code> .
------	--

`KafkaMicroBatchInputPartitionReader` initializes the [internal properties](#).

## next Method

next(): Boolean
-----------------

Note	<code>next</code> is part of the <code>InputPartitionReader</code> contract to proceed to next record if available ( <code>true</code> ).
------	---

`next` checks whether the `KafkaDataConsumer` should [poll records](#) or [not](#) (i.e. `nextOffset` is smaller than the `untilOffset` of the `KafkaOffsetRange`).

## next Method — KafkaDataConsumer Polls Records

If so, `next` requests the `KafkaDataConsumer` to get (`poll`) records in the range of `nextOffset` and the `untilOffset` (of the `KafkaOffsetRange`) with the given `pollTimeoutMs` and `failOnDataLoss`.

With a new record, `next` requests the [KafkaRecordToUnsafeRowConverter](#) to convert (`toUnsafeRow`) the record to be the `next UnsafeRow`. `next` sets the `nextOffset` as the offset of the record incremented. `next` returns `true`.

With no new record, `next` simply returns `false`.

## next Method — No Polling

If the `nextOffset` is equal or larger than the `untilOffset` (of the [KafkaOffsetRange](#)), `next` simply returns `false`.

## Closing (Releasing KafkaDataConsumer) — close Method

```
close(): Unit
```

Note	<code>close</code> is part of the Java Closeable contract to release resources.
------	---

`close` simply requests the [KafkaDataConsumer](#) to `release`.

## resolveRange Internal Method

```
resolveRange(  
    range: KafkaOffsetRange): KafkaOffsetRange
```

`resolveRange` ...FIXME

Note	<code>resolveRange</code> is used exclusively when <code>KafkaMicroBatchInputPartitionReader</code> is created (and initializes the <a href="#">KafkaOffsetRange</a> internal property).
------	--

## Internal Properties

Name	Description
consumer	<code>KafkaDataConsumer</code> for the partition (per <code>KafkaOffsetRange</code> ) Used in <code>next</code> , <code>close</code> , and <code>resolveRange</code>
converter	<code>KafkaRecordToUnsafeRowConverter</code>
nextOffset	<code>Next offset</code>
nextRow	<code>Next UnsafeRow</code>
rangeToRead	<code>KafkaOffsetRange</code>

# KafkaSourceInitialOffsetWriter

`KafkaSourceInitialOffsetWriter` is a Hadoop DFS-based metadata storage for `KafkaSourceOffsets`.

`KafkaSourceInitialOffsetWriter` is created exclusively when `KafkaMicroBatchReader` is requested to `getOrCreateInitialPartitionOffsets`.

`KafkaSourceInitialOffsetWriter` uses `1` for the version.

## Creating KafkaSourceInitialOffsetWriter Instance

`KafkaSourceInitialOffsetWriter` takes the following to be created:

- `SparkSession`
- Path of the metadata log directory

## Deserializing Metadata (Reading Metadata from Serialized Format) — `deserialize` Method

```
deserialize(  
    in: InputStream): KafkaSourceOffset
```

Note

`deserialize` is part of the [HDFSMetadataLog Contract](#) to deserialize metadata (reading metadata from a serialized format)

`deserialize` ...FIXME

# KafkaContinuousReader — ContinuousReader for Kafka Data Source in Continuous Stream Processing

`KafkaContinuousReader` is a [ContinuousReader](#) for [Kafka Data Source](#) in [Continuous Stream Processing](#).

`KafkaContinuousReader` is [created](#) exclusively when `KafkaSourceProvider` is requested to [create a ContinuousReader](#).

`KafkaContinuousReader` uses `kafkaConsumer.pollTimeoutMs` configuration parameter (default: `512`) for [KafkaContinuousInputPartitions](#) when requested to [planInputPartitions](#).

<b>Tip</b>	<p>Enable <code>INFO</code> or <code>WARN</code> logging levels for <code>org.apache.spark.sql.kafka010.KafkaContinuousReader</code> to see what happens inside.</p> <p>Add the following line to <code>conf/log4j.properties</code> :</p> <pre>log4j.logger.org.apache.spark.sql.kafka010.KafkaContinuousReader=INFO</pre> <p>Refer to <a href="#">Logging</a>.</p>
------------	--

## Creating KafkaContinuousReader Instance

`KafkaContinuousReader` takes the following to be created:

- [KafkaOffsetReader](#)
- Kafka parameters (as `java.util.Map[String, Object]`)
- Source options (as `Map[String, String]`)
- Metadata path
- [Initial offsets](#)
- `failOnDataLoss` flag

## Plan Input Partitions — `planInputPartitions` Method

```
planInputPartitions(): java.util.List[InputPartition[InternalRow]]
```

**Note**

`planInputPartitions` is part of the `DataSourceReader` contract in Spark SQL for the number of `InputPartitions` to use as RDD partitions (when `DataSourceV2ScanExec` physical operator is requested for the partitions of the input RDD).

`planInputPartitions` ...FIXME

## **setStartOffset Method**

```
setStartOffset(  
    start: Optional[Offset]): Unit
```

**Note**

`setStartoffset` is part of the [ContinuousReader Contract](#) to...FIXME.

`setStartOffset` ...FIXME

## **deserializeOffset Method**

```
deserializeOffset(  
    json: String): Offset
```

**Note**

`deserializeOffset` is part of the [ContinuousReader Contract](#) to...FIXME.

`deserializeOffset` ...FIXME

## **mergeOffsets Method**

```
mergeOffsets(  
    offsets: Array[PartitionOffset]): Offset
```

**Note**

`mergeOffsets` is part of the [ContinuousReader Contract](#) to...FIXME.

`mergeOffsets` ...FIXME

# KafkaContinuousInputPartition

KafkaContinuousInputPartition is...FIXME

# TextSocketSourceProvider

`TextSocketSourceProvider` is a [StreamSourceProvider](#) for [TextSocketSource](#) that read records from `host` and `port`.

`TextSocketSourceProvider` is a [DataSourceRegister](#), too.

The short name of the data source is `socket`.

It requires two mandatory options (that you can set using `option` method):

1. `host` which is the host name.
2. `port` which is the port number. It must be an integer.

`TextSocketSourceProvider` also supports [includeTimestamp](#) option that is a boolean flag that you can use to include timestamps in the schema.

## includeTimestamp Option

Caution	FIXME
---------	-------

## createSource

`createSource` grabs the two mandatory options — `host` and `port` — and returns an [TextSocketSource](#).

## sourceSchema

`sourceSchema` returns `textSocket` as the name of the source and the schema that can be one of the two available schemas:

1. `SCHEMA_REGULAR` (default) which is a schema with a single `value` field of String type.
2. `SCHEMA_TIMESTAMP` when `includeTimestamp` flag option is set. It is not, i.e. `false`, by default. The schema are `value` field of `StringType` type and `timestamp` field of `TimestampType` type of format `yyyy-MM-dd HH:mm:ss`.

Tip	Read about <a href="#">schema</a> .
-----	-------------------------------------

Internally, it starts by printing out the following WARN message to the logs:

```
WARN TextSocketSourceProvider: The socket source should not be used for production app  
lications! It does not support recovery and stores state indefinitely.
```

It then checks whether `host` and `port` parameters are defined and if not it throws a `AnalysisException` :

```
Set a host to read from with option("host", ...).
```

# TextSocketSource

`TextSocketSource` is a [streaming source](#) that reads lines from a socket at the `host` and `port` (defined by parameters).

It uses `lines` internal in-memory buffer to keep all of the lines that were read from a socket forever.

Caution	This source is <b>not</b> for production use due to design constraints, e.g. infinite in-memory collection of lines read and no fault recovery.  It is designed only for tutorials and debugging.
---------	---

```

import org.apache.spark.sql.SparkSession
val spark: SparkSession = SparkSession.builder.getOrCreate()

// Connect to localhost:9999
// You can use "nc -lk 9999" for demos
val textSocket = spark.
  readStream.
  format("socket").
  option("host", "localhost").
  option("port", 9999).
  load

import org.apache.spark.sql.Dataset
val lines: Dataset[String] = textSocket.as[String].map(_.toUpperCase)

val query = lines.writeStream.format("console").start

// Start typing the lines in nc session
// They will appear UPPERCASE in the terminal

-----
Batch: 0
-----
+---+
|   value|
+---+
|UPPERCASE|
+---+

scala> query.explain
== Physical Plan ==
*SerializeFromObject [staticinvoke(class org.apache.spark.unsafe.types.UTF8String, StringType, fromString, input[0, java.lang.String, true], true) AS value#21]
+- *MapElements <function1>, obj#20: java.lang.String
  +- *DeserializeToObject value#43.toString, obj#19: java.lang.String
    +- LocalTableScan [value#43]

scala> query.stop

```

## lines Internal Buffer

```
lines: ArrayBuffer[(String, Timestamp)]
```

`lines` is the internal buffer of all the lines `TextSocketSource` read from the socket.

## Maximum Available Offset (getOffset method)

**Note**

`getOffset` is a part of the [Streaming Source Contract](#).

`TextSocketSource`'s offset can either be none or `Longoffset` of the number of lines in the internal `lines` buffer.

## Schema (schema method)

`TextSocketSource` supports two [schemas](#):

1. A single `value` field of String type.
2. `value` field of `StringType` type and `timestamp` field of `TimestampType` type of format `yyyy-MM-dd HH:mm:ss`.

**Tip**

Refer to [sourceSchema](#) for `TextSocketSourceProvider`.

## Creating TextSocketSource Instance

```
TextSocketSource(  
    host: String,  
    port: Int,  
    includeTimestamp: Boolean,  
    sqlContext: SQLContext)
```

When `TextSocketSource` is created (see [TextSocketSourceProvider](#)), it gets 4 parameters passed in:

1. `host`
2. `port`
3. [includeTimestamp](#) flag
4. [SQLContext](#)

**Caution**

It appears that the source did not get "renewed" to use [SparkSession](#) instead.

It opens a socket at given `host` and `port` parameters and reads a buffering character-input stream using the default charset and the default-sized input buffer (of `8192` bytes) line by line.

**Caution**

**FIXME** Review Java's `Charset.defaultCharset()`

It starts a `readThread` daemon thread (called `TextSocketSource(host, port)` ) to read lines from the socket. The lines are added to the internal `lines` buffer.

## Stopping TextSocketSource (stop method)

When stopped, `TextSocketSource` closes the socket connection.

# RateSourceProvider

`RateSourceProvider` is a [StreamSourceProvider](#) for [RateStreamSource](#) (that acts as the source for **rate** format).

Note	<code>RateSourceProvider</code> is also a <code>DataSourceRegister</code> .
------	---

The short name of the data source is **rate**.

# RateStreamSource

`RateStreamSource` is a [streaming source](#) that generates [consecutive numbers](#) with [timestamp](#) that can be useful for testing and PoCs.

`RateStreamSource` is created for **rate** format (that is registered by [RateSourceProvider](#)).

```
val rates = spark
  .readStream
  .format("rate") // <-- use RateStreamSource
  .option("rowsPerSecond", 1)
  .load
```

Table 1. RateStreamSource's Options

Name	Default Value	Description
<code>numPartitions</code>	(default parallelism)	Number of partitions to use
<code>rampUpTime</code>	0 (seconds)	
<code>rowsPerSecond</code>	1	Number of rows to generate per second (has to be greater than 0 )

`RateStreamSource` uses a predefined schema that cannot be changed.

```
val schema = rates.schema
scala> println(schema.treeString)
root
| -- timestamp: timestamp (nullable = true)
| -- value: long (nullable = true)
```

Table 2. RateStreamSource's Dataset Schema (in the positional order)

Name	Type
<code>timestamp</code>	<code>TimestampType</code>
<code>value</code>	<code>LongType</code>

Table 3. RateStreamSource's Internal Registries and Counters

Name	Description
clock	
lastTimeMs	
maxSeconds	
startTimeMs	

Tip	<p>Enable <code>INFO</code> or <code>DEBUG</code> logging levels for <code>org.apache.spark.sql.execution.streaming.RateStreamSource</code> to see what happens inside.</p> <p>Add the following line to <code>conf/log4j.properties</code> :</p> <pre>log4j.logger.org.apache.spark.sql.execution.streaming.RateStreamSource=DEBUG</pre> <p>Refer to <a href="#">Logging</a>.</p>

## Getting Maximum Available Offsets — `getOffset` Method

```
getOffset: Option[Offset]
```

Note	<code>getOffset</code> is a part of the <a href="#">Source Contract</a> .
Caution	FIXME

## Generating DataFrame for Streaming Batch — `getBatch` Method

```
getBatch(start: Option[Offset], end: Offset): DataFrame
```

Note	<code>getBatch</code> is a part of <a href="#">Source Contract</a> .
------	--

Internally, `getBatch` calculates the seconds to start from and end at (from the input `start` and `end` offsets) or assumes `0`.

`getBatch` then calculates the values to generate for the start and end seconds.

You should see the following DEBUG message in the logs:

```
DEBUG RateStreamSource: startSeconds: [startSeconds], endSeconds: [endSeconds], rangeStart: [rangeStart], rangeEnd: [rangeEnd]
```

If the start and end ranges are equal, `getBatch` creates an empty `DataFrame` (with the [schema](#)) and returns.

Otherwise, when the ranges are different, `getBatch` creates a `DataFrame` using `SparkContext.range` operator (for the start and end ranges and [numPartitions](#) partitions).

## Creating RateStreamSource Instance

`RateStreamSource` takes the following when created:

- `SQLContext`
- Path to the metadata
- Rows per second
- RampUp time in seconds
- Number of partitions
- Flag to whether to use `ManualClock` (`true`) or `SystemClock` (`false`)

`RateStreamSource` initializes the [internal registries and counters](#).

# RateStreamMicroBatchReader

RateStreamMicroBatchReader is...FIXME

# ConsoleSinkProvider

`ConsoleSinkProvider` is a `DataSourcev2` with `StreamWriterSupport` for **console** data source format.

Tip	Read up on <a href="#">DataSourceV2 Contract</a> in <a href="#">The Internals of Spark SQL</a> book.
-----	--

`ConsoleSinkProvider` is a `DataSourceRegister` and registers itself as the **console** data source format.

```
import org.apache.spark.sql.streaming.Trigger
val q = spark
  .readStream
  .format("rate")
  .load
  .writeStream
  .format("console") // <-- requests ConsoleSinkProvider for a sink
  .trigger(Trigger.Once)
  .start
scala> println(q.lastProgress.sink)
{
  "description" : "org.apache.spark.sql.execution.streaming.ConsoleSinkProvider@2392cf
b1"
}
```

When requested for a `StreamWriter`, `ConsoleSinkProvider` simply creates a `ConsoleWriter` (with the given schema and options).

`ConsoleSinkProvider` is a [CreatableRelationProvider](#).

Tip	Read up on <a href="#">CreatableRelationProvider</a> in <a href="#">The Internals of Spark SQL</a> book.
-----	--

## createRelation Method

```
createRelation(
  sqlContext: SQLContext,
  mode: SaveMode,
  parameters: Map[String, String],
  data: DataFrame): BaseRelation
```

Note	<code>createRelation</code> is part of the <code>CreatableRelationProvider</code> Contract to support writing a structured query (a <code>DataFrame</code> ) per save mode.
------	---

`createRelation ...FIXME`



# ConsoleWriter

ConsoleWriter is a [StreamWriter](#) for **console** data source format.

# ForeachWriterProvider

ForeachWriterProvider is...FIXME

# ForeachWriter

`ForeachWriter` is the [contract](#) for a **foreach writer** that is a [streaming format](#) that controls streaming writes.

Note	<code>ForeachWriter</code> is set using <a href="#">foreach</a> operator.
------	---

```
val foreachWriter = new ForeachWriter[String] { ... }
streamingQuery.
writeStream.
foreach(foreachWriter).
start
```

## ForeachWriter Contract

```
package org.apache.spark.sql

abstract class ForeachWriter[T] {
  def open(partitionId: Long, version: Long): Boolean
  def process(value: T): Unit
  def close(errorOrNull: Throwable): Unit
}
```

Table 1. ForeachWriter Contract

Method	Description
<code>open</code>	Used when...
<code>process</code>	Used when...
<code>close</code>	Used when...

# ForeachSink

`ForeachSink` is a typed [streaming sink](#) that passes rows (of the type `T`) to [ForeachWriter](#) (one record at a time per partition).

Note	<code>ForeachSink</code> is assigned a <code>ForeachWriter</code> when <code>DataStreamWriter</code> is <a href="#">started</a> .
------	---

`ForeachSink` is used exclusively in [foreach](#) operator.

```
val records = spark.  
  readStream  
  format("text").  
  load("server-logs/*.out").  
  as[String]  
  
import org.apache.spark.sql.ForeachWriter  
val writer = new ForeachWriter[String] {  
  override def open(partitionId: Long, version: Long) = true  
  override def process(value: String) = println(value)  
  override def close(errorOrNull: Throwable) = {}  
}  
  
records.writeStream  
  .queryName("server-logs processor")  
  .foreach(writer)  
  .start
```

Internally, `addBatch` (the only method from the [Sink Contract](#)) takes records from the input [DataFrame](#) (as `data`), transforms them to expected type `T` (of this `ForeachSink`) and (now as a [Dataset](#)) processes each partition.

addBatch(batchId: Long, data: DataFrame): Unit
--

`addBatch` then opens the constructor's [ForeachWriter](#) (for the [current partition](#) and the input batch) and passes the records to process (one at a time per partition).

Caution	FIXME Why does Spark track whether the writer failed or not? Why couldn't it <code>finally</code> and do <code>close</code> ?
---------	---

Caution	FIXME Can we have a constant for "foreach" for <code>source</code> in <code>DataStreamWriter</code> ?
---------	---



# ForeachBatchSink

`ForeachBatchSink` is a [streaming sink](#) that is used for the `DataStreamWriter.foreachBatch` streaming operator.

`ForeachBatchSink` is created exclusively when `DataStreamWriter` is requested to [start execution of the streaming query](#) (with the `foreachBatch` source).

`ForeachBatchSink` uses **ForeachBatchSink** name.

```
import org.apache.spark.sql.Dataset
val q = spark.readStream
  .format("rate")
  .load
  .writeStream
  .foreachBatch { (output: Dataset[_], batchId: Long) => // <-- creates a ForeachBatch
    sink
    println(s"Batch ID: $batchId")
    output.show
  }
  .start
// q.stop

scala> println(q.lastProgress.sink.description)
ForeachBatchSink
```

Note

`ForeachBatchSink` was added in Spark 2.4.0 as part of [SPARK-24565 Add API for in Structured Streaming for exposing output rows of each microbatch as a DataFrame](#).

## Creating ForeachBatchSink Instance

`ForeachBatchSink` takes the following when created:

- Batch writer (`(Dataset[T], Long) ⇒ Unit`)
- Encoder (`ExpressionEncoder[T]`)

## Adding Batch — `addBatch` Method

```
addBatch(batchId: Long, data: DataFrame): Unit
```

Note

`addBatch` is a part of [Sink Contract](#) to "add" a batch of data to the sink.

addBatch ...**FIXME**

# Memory Data Source

**Memory Data Source** is made up of the following two base implementations to support the older DataSource API V1 and the modern DataSource API V2:

- [MemoryStreamBase](#)
- [MemorySinkBase](#)

Memory data source supports [Micro-Batch](#) and [Continuous](#) stream processing modes.

Stream Processing	Source	Sink
<a href="#">Micro-Batch</a>	<a href="#">MemoryStream</a>	<a href="#">MemorySink</a>
<a href="#">Continuous</a>	<a href="#">ContinuousMemoryStream</a>	<a href="#">MemorySinkV2</a>

Caution	<p>Memory Data Source is <b>not</b> for production use due to design constraints, e.g. infinite in-memory collection of lines read and no fault recovery.</p> <p><code>MemoryStream</code> is designed primarily for unit tests, tutorials and debugging.</p>
---------	---

## Memory Sink

Memory sink requires that a streaming query has a name (defined using [DataStreamWriter.queryName](#) or `queryName` option).

Memory sink may optionally define checkpoint location using `checkpointLocation` option that is used to recover from for [Complete](#) output mode only.

## Memory Sink and CreateViewCommand

When a streaming query with `memory` sink is [started](#), [DataStreamWriter](#) uses `Dataset.createOrReplaceTempView` operator to create or replace a local temporary view with the name of the query (which is required).



## Details for Query 0

Submitted Time: 2019/10/12 17:45:37

Duration: 2 ms

[Execute CreateViewCommand](#)

▼ Details

```
== Parsed Logical Plan ==
CreateViewCommand `StreamingAggregationApp`, false, true, LocalTempView
  +- MemoryPlan MemorySink, [sliding_window#24, batches#25, values#26]

== Analyzed Logical Plan ==
CreateViewCommand `StreamingAggregationApp`, false, true, LocalTempView
  +- MemoryPlan MemorySink, [sliding_window#24, batches#25, values#26]

== Optimized Logical Plan ==
CreateViewCommand `StreamingAggregationApp`, false, true, LocalTempView
  +- MemoryPlan MemorySink, [sliding_window#24, batches#25, values#26]

== Physical Plan ==
Execute CreateViewCommand
  +- CreateViewCommand `StreamingAggregationApp`, false, true, LocalTempView
    +- MemoryPlan MemorySink, [sliding_window#24, batches#25, values#26]
```

Figure 1. Memory Sink and CreateViewCommand

## Examples

### Memory Source in Micro-Batch Stream Processing

```

val spark: SparkSession = ???

implicit val ctx = spark.sqlContext

import org.apache.spark.sql.execution.streaming.MemoryStream
// It uses two implicits: Encoder[Int] and SQLContext
val intsIn = MemoryStream[Int]

val ints = intsIn.toDF
  .withColumn("t", current_timestamp())
  .withWatermark("t", "5 minutes")
  .groupBy(window($"t", "5 minutes") as "window")
  .agg(count("*") as "total")

import org.apache.spark.sql.streaming.{OutputMode, Trigger}
import scala.concurrent.duration._
val totalsOver5mins = ints.
  writeStream.
  format("memory").
  queryName("totalsOver5mins").
  outputMode(OutputMode.Append).
  trigger(Trigger.ProcessingTime(10.seconds)).
  start

val zeroOffset = intsIn.addData(0, 1, 2)
totalsOver5mins.processAllAvailable()
spark.table("totalsOver5mins").show

scala> intsOut.show
+---+
|value|
+---+
|   0|
|   1|
|   2|
+---+

memoryQuery.stop()

```

## Memory Sink in Micro-Batch Stream Processing

```
val queryName = "memoryDemo"
val sq = spark
  .readStream
  .format("rate")
  .load
  .writeStream
  .format("memory")
  .queryName(queryName)
  .start

// The name of the streaming query is an in-memory table
val showAll = sql(s"select * from $queryName")
scala> showAll.show(truncate = false)
+-----+-----+
|timestamp          |value|
+-----+-----+
|2019-10-10 15:19:16.431|42   |
|2019-10-10 15:19:17.431|43   |
+-----+-----+

import org.apache.spark.sql.streaming.StreamingQuery
assert(sq.isInstanceOf[StreamingQuery])

import org.apache.spark.sql.execution.streaming.StreamingQueryWrapper
val se = sq.asInstanceOf[StreamingQueryWrapper].streamingQuery

import org.apache.spark.sql.execution.streaming.MemorySink
val sink = se.sink.asInstanceOf[MemorySink]

assert(sink.toString == "MemorySink")

sink.clear()
```

# MemoryStream — Streaming Reader for Micro-Batch Stream Processing

`MemoryStream` is a concrete [streaming source](#) of [memory data source](#) that supports [reading](#) in [Micro-Batch Stream Processing](#).

Tip

Enable `ALL` logging level for `org.apache.spark.sql.execution.streaming.MemoryStream` logger to see what happens inside.

Add the following line to `conf/log4j.properties` :

```
log4j.logger.org.apache.spark.sql.execution.streaming.MemoryStream=ALL
```

Refer to [Logging](#).

## Creating MemoryStream Instance

`MemoryStream` takes the following to be created:

- `ID`
- `SQLContext`

`MemoryStream` initializes the [internal properties](#).

## Creating MemoryStream Instance — `apply` Object Factory

```
apply[A : Encoder](
  implicit sqlContext: SQLContext): MemoryStream[A]
```

`apply` uses an `memoryStreamId` internal counter to [create a new `MemoryStream`](#) with a unique `ID` and the implicit `sqlContext`.

## Adding Data to Source — `addData` Method

```
addData(
  data: TraversableOnce[A]): Offset
```

`addData` adds the given `data` to the `batches` internal registry.

Internally, `addData` prints out the following DEBUG message to the logs:

```
Adding: [data]
```

In the end, `addData` increments the `current offset` and adds the data to the `batches` internal registry.

## Generating Next Streaming Batch — `getBatch` Method

Note

`getBatch` is a part of [Streaming Source contract](#).

When executed, `getBatch` uses the internal `batches` collection to return requested offsets.

You should see the following DEBUG message in the logs:

```
DEBUG MemoryStream: MemoryBatch [[startOrdinal], [endOrdinal]]: [newBlocks]
```

## Logical Plan — `logicalPlan` Internal Property

```
logicalPlan: LogicalPlan
```

Note

`logicalPlan` is part of the [MemoryStreamBase Contract](#) for the logical query plan of the memory stream.

`logicalPlan` is simply a [StreamingExecutionRelation](#) (for this memory source and the [attributes](#)).

`MemoryStream` uses [StreamingExecutionRelation](#) logical plan to build [Datasets](#) or [DataFrames](#) when requested.

```
scala> val ints = MemoryStream[Int]
ints: org.apache.spark.sql.execution.streaming.MemoryStream[Int] = MemoryStream[value#13]

scala> ints.toDS.queryExecution.logical.isStreaming
res14: Boolean = true

scala> ints.toDS.queryExecution.logical
res15: org.apache.spark.sql.catalyst.plans.logical.LogicalPlan = MemoryStream[value#13]
```

## Schema (schema method)

`MemoryStream` works with the data of the [schema](#) as described by the [Encoder](#) (of the `Dataset` ).

## Textual Representation — `toString` Method

```
toString: String
```

Note

`toString` is part of the [java.lang.Object](#) contract for the string representation of the object.

`toString` uses the [output schema](#) to return the following textual representation:

```
MemoryStream[[output]]
```

## Plan Input Partitions — `planInputPartitions` Method

```
planInputPartitions(): java.util.List[InputPartition[InternalRow]]
```

Note

`planInputPartitions` is part of the `DataSourceReader` contract in Spark SQL for the number of `InputPartitions` to use as RDD partitions (when `DataSourceV2ScanExec` physical operator is requested for the partitions of the input RDD).

`planInputPartitions` ...FIXME

`planInputPartitions` prints out a DEBUG message to the logs with the [generateDebugString](#) (with the batches after the [last committed offset](#)).

`planInputPartitions` ...FIXME

## generateDebugString Internal Method

```
generateDebugString(
  rows: Seq[UnsafeRow],
  startOrdinal: Int,
  endOrdinal: Int): String
```

`generateDebugString` resolves and binds the [encoder](#) for the data.

In the end, `generateDebugString` returns the following string:

```
MemoryBatch [[startOrdinal], [endOrdinal]]: [rows]
```

**Note**

`generateDebugString` is used exclusively when `MemoryStream` is requested to [planInputPartitions](#).

## Internal Properties

Name	Description
batches	Batch data ( <code>ListBuffer[Array[UnsafeRow]]</code> )
currentOffset	Current offset (as <code>LongOffset</code> )
lastOffsetCommitted	Last committed offset (as <code>LongOffset</code> )
output	Output schema ( <code>Seq[Attribute]</code> ) of the logical query plan Used exclusively for <code>toString</code>

# ContinuousMemoryStream

ContinuousMemoryStream is...FIXME

# MemorySink

`MemorySink` is a [streaming sink](#) that stores batches (records) in memory.

`MemorySink` is intended only for testing or demos.

`MemorySink` is used for `memory` format and requires a query name (by `queryName` method or `queryName` option).

Note

`MemorySink` was introduced in the pull request for [\[SPARK-14288\]\[SQL\]](#) [Memory Sink for streaming](#).

Use `toDebugString` to see the batches.

Its aim is to allow users to test streaming applications in the Spark shell or other local tests.

You can set `checkpointLocation` using `option` method or it will be set to `spark.sql.streaming.checkpointLocation` property.

If `spark.sql.streaming.checkpointLocation` is set, the code uses `$location/$queryName` directory.

Finally, when no `spark.sql.streaming.checkpointLocation` is set, a temporary directory `memory.stream` under `java.io.tmpdir` is used with `offsets` subdirectory inside.

Note

The directory is cleaned up at shutdown using `ShutdownHookManager.registerShutdownDeleteDir`.

It creates `MemorySink` instance based on the schema of the DataFrame it operates on.

It creates a new DataFrame using `MemoryPlan` with `MemorySink` instance created earlier and registers it as a temporary table (using `DataFrame.registerTempTable` method).

Note

At this point you can query the table as if it were a regular non-streaming table using `sql` method.

A new `StreamingQuery` is started (using `StreamingQueryManager.startQuery`) and returned.

Enable `ALL` logging level for `org.apache.spark.sql.execution.streaming.MemorySink` logger to see what happens inside.

Add the following line to `conf/log4j.properties`:

Tip

```
log4j.logger.org.apache.spark.sql.execution.streaming.MemorySink=ALL
```

Refer to [Logging](#).

## Creating MemorySink Instance

`MemorySink` takes the following to be created:

- Output schema
- [OutputMode](#)

`MemorySink` initializes the `batches` internal property.

## In-Memory Buffer of Streaming Batches — `batches` Internal Property

```
batches: ArrayBuffer[AddedData]
```

`batches` holds data from streaming batches that have been [added](#) (*written*) to this sink.

For [Append](#) and [Update](#) output modes, `batches` holds rows from all batches.

For [Complete](#) output mode, `batches` holds rows from the last batch only.

`batches` can be cleared (*emptied*) using [clear](#).

## Adding Batch of Data to Sink — `addBatch` Method

```
addBatch(  
    batchId: Long,  
    data: DataFrame): Unit
```

Note `addBatch` is part of the [Sink Contract](#) to "add" a batch of data to the sink.

`addBatch` branches off based on whether the given `batchId` has already been [committed](#) or [not](#).

A batch ID is considered **committed** when the given batch ID is greater than the [latest batch ID](#) (if available).

## Batch Not Committed

With the `batchId` not committed, `addBatch` prints out the following DEBUG message to the logs:

```
Committing batch [batchId] to [this]
```

`addBatch` collects records from the given `data`.

Note	<code>addBatch</code> uses <code>Dataset.collect</code> operator to collect records.
------	--

For [Append](#) and [Update](#) output modes, `addBatch` adds the data (as a `AddedData`) to the [batches](#) internal registry.

For [Complete](#) output mode, `addBatch` clears the [batches](#) internal registry first before adding the data (as a `AddedData`).

For any other output mode, `addBatch` reports an `IllegalArgumentException`:

```
Output mode [outputMode] is not supported by MemorySink
```

## Batch Committed

With the `batchId` committed, `addBatch` simply prints out the following DEBUG message to the logs and returns.

```
Skipping already committed batch: [batchId]
```

## Clearing Up Internal Batch Buffer— `clear` Method

```
clear(): Unit
```

`clear` simply removes (*clears*) all data from the [batches](#) internal registry.

Note	<code>clear</code> is used exclusively in tests.
------	--



# MemorySinkV2 — Writable Streaming Sink for Continuous Stream Processing

`MemorySinkV2` is a `DataSourceV2` with [StreamWriterSupport](#) for **memory** data source format in [Continuous Stream Processing](#).

Tip

Read up on [DataSourceV2 Contract](#) in [The Internals of Spark SQL](#) book.

`MemorySinkV2` is a custom [MemorySinkBase](#).

When requested for a [StreamWriter](#), `MemorySinkV2` simply creates a [MemoryStreamWriter](#).

# MemoryStreamWriter

MemoryStreamWriter is...FIXME

# MemoryStreamBase Contract — Base Contract for Memory Sources

`MemoryStreamBase` is the [base](#) of the [BaseStreamingSource contract](#) for [memory sources](#) that can [add data](#).

Table 1. MemoryStreamBase Contract

Method	Description
<code>addData</code>	<code>addData(   data: TraversableOnce[A]): Offset</code>
<code>logicalPlan</code>	<code>logicalPlan: LogicalPlan</code>

Table 2. MemoryStreamBases

MemoryStreamBase	Description
<a href="#">ContinuousMemoryStream</a>	
<a href="#">MemoryStream</a>	<a href="#">MicroBatchReader</a> for <a href="#">Micro-Batch Stream Processing</a>

## Creating MemoryStreamBase Instance

`MemoryStreamBase` takes the following to be created:

- `SQLContext`

Note

`MemoryStreamBase` is a Scala abstract class and cannot be [created](#) directly. It is created indirectly for the [concrete MemoryStreamBases](#).

## Creating Streaming Dataset — `toDS` Method

```
toDS(): Dataset[A]
```

`toDS` simply creates a `Dataset` (for the `sqlContext` and the `logicalPlan`)

## Creating Streaming DataFrame — `toDF` Method

```
toDF(): DataFrame
```

`toDF` simply creates a `Dataset` of rows (for the `sqlContext` and the `logicalPlan`)

## Internal Properties

Name	Description
<code>attributes</code>	Schema attributes of the <code>encoder</code> ( <code>Seq[AttributeReference]</code> )  Used when...FIXME
<code>encoder</code>	Spark SQL's <code>ExpressionEncoder</code> for the data  Used when...FIXME

# MemorySinkBase Contract — Base Contract for Memory Sinks

`MemorySinkBase` is the [extension](#) of the [BaseStreamingSink contract](#) for memory sinks that manage [all data](#) in memory.

Table 1. MemorySinkBase Contract

Method	Description
<code>allData</code>	<code>allData: Seq[Row]</code>
<code>dataSinceBatch</code>	<code>dataSinceBatch(   sinceBatchId: Long): Seq[Row]</code>
<code>latestBatchData</code>	<code>latestBatchData: Seq[Row]</code>
<code>latestBatchId</code>	<code>latestBatchId: Option[Long]</code>

Table 2. MemorySinkBases

MemorySinkBase	Description
<a href="#">MemorySink</a>	<a href="#">Streaming sink for Micro-Batch Stream Processing</a> (based on Data Source API V1)
<a href="#">MemorySinkV2</a>	<a href="#">Writable streaming sink for Continuous Stream Processing</a> (based on Data Source API V2)

# Offsets and Metadata Checkpointing

A streaming query can be started from scratch or from checkpoint (that gives fault-tolerance as the state is preserved even when a failure happens).

`StreamExecution` use **checkpoint location** to resume stream processing and get **start offsets** to start query processing from.

`StreamExecution` resumes (populates the start offsets) from the latest checkpointed offsets from the [Write-Ahead Log \(WAL\) of Offsets](#) that may have already been processed (and, if so, committed to the [Offset Commit Log](#)).

- [Hadoop DFS-based metadata storage](#) of `OffsetSeqs`
- `OffsetSeq` and [StreamProgress](#)
- [StreamProgress](#) and [StreamExecutions](#) ([committed](#) and [available offsets](#))

## Micro-Batch Stream Processing

In [Micro-Batch Stream Processing](#), the [available offsets](#) registry is populated with the [latest offsets](#) from the [Write-Ahead Log \(WAL\)](#) when `MicroBatchExecution` stream processing engine is requested to [populate start offsets from checkpoint](#) (if available) when `MicroBatchExecution` is requested to run an activated streaming query ([before the first "zero" micro-batch](#)).

The [available offsets](#) are then [added](#) to the [committed offsets](#) when the latest batch ID available (as described above) is exactly the [latest batch ID](#) committed to the [Offset Commit Log](#) when `MicroBatchExecution` stream processing engine is requested to [populate start offsets from checkpoint](#).

When a streaming query is started from scratch (with no checkpoint that has offsets in the [Offset Write-Ahead Log](#)), `MicroBatchExecution` prints out the following INFO message:

```
Starting new streaming query.
```

When a streaming query is resumed (restarted) from a checkpoint with offsets in the [Offset Write-Ahead Log](#), `MicroBatchExecution` prints out the following INFO message:

```
Resuming at batch [currentBatchId] with committed offsets
[committedOffsets] and available offsets [availableOffsets]
```

Every time `MicroBatchExecution` is requested to check whether a new data is available (in any of the streaming sources)...FIXME

When `MicroBatchExecution` is requested to construct the next streaming micro-batch (when `MicroBatchExecution` requested to run the activated streaming query), every streaming source is requested for the latest offset available that are added to the `availableOffsets` registry. Streaming sources report some offsets or none at all (if this source has never received any data). Streaming sources with no data are excluded (*filtered out*).

`MicroBatchExecution` prints out the following TRACE message to the logs:

```
noDataBatchesEnabled = [noDataBatchesEnabled],  
lastExecutionRequiresAnotherBatch =  
[lastExecutionRequiresAnotherBatch], isNewDataAvailable =  
[isNewDataAvailable], shouldConstructNextBatch =  
[shouldConstructNextBatch]
```

With `shouldConstructNextBatch` internal flag enabled, `MicroBatchExecution` commits (adds) the available offsets for the batch to the Write-Ahead Log (WAL) and prints out the following INFO message to the logs:

```
Committed offsets for batch [currentBatchId]. Metadata  
[offsetSeqMetadata]
```

When running a single streaming micro-batch, `MicroBatchExecution` requests every Source and MicroBatchReader (in the `availableOffsets` registry) for unprocessed data (that has not been committed yet and so considered unprocessed).

In the end (of running a single streaming micro-batch), `MicroBatchExecution` commits (adds) the available offsets (to the `committedOffsets` registry) so they are considered processed already.

`MicroBatchExecution` prints out the following DEBUG message to the logs:

```
Completed batch [currentBatchId]
```

## Limitations (Assumptions)

It is assumed that the order of streaming sources in a streaming query matches the order of the offsets of OffsetSeq (in `offsetLog`) and `availableOffsets`.

In other words, a streaming query can be modified and then restarted from a checkpoint (to maintain stream processing state) only when the number of streaming sources and their order are preserved across restarts.

# MetadataLog Contract — Metadata Storage

`MetadataLog` is the abstraction of metadata storage that can persist, retrieve, and remove metadata (of type `T`).

Table 1. MetadataLog Contract

Method	Description
<code>add</code>	<pre>add(     batchId: Long,     metadata: T): Boolean</pre> <p>Persists (<i>adds</i>) metadata of a streaming batch</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>KafkaMicroBatchReader</code> is requested to <code>getOrCreateInitialPartitionOffsets</code></li> <li>• <code>KafkaSource</code> is requested for the <code>initialPartitionOffsets</code></li> <li>• <code>CompactibleFileStreamLog</code> is requested for the <code>store metadata of a streaming batch</code> and to <code>compact</code></li> <li>• <code>FileStreamSource</code> is requested to <code>fetchMaxOffset</code></li> <li>• <code>FileStreamSourceLog</code> is requested to <code>store (add) metadata of a streaming batch</code></li> <li>• <code>ManifestFileCommitProtocol</code> is requested to <code>commitJob</code></li> <li>• <code>MicroBatchExecution</code> stream execution engine is requested to <code>construct a next streaming micro-batch</code> and <code>run a single streaming micro-batch</code></li> <li>• <code>ContinuousExecution</code> stream execution engine is requested to <code>addOffset</code> and <code>commit an epoch</code></li> <li>• <code>RateStreamMicroBatchReader</code> is created (<code>creationTimeMs</code>)</li> </ul>
<code>get</code>	<pre>get(     batchId: Long): Option[T] get(     startId: Option[Long],     endId: Option[Long]): Array[(Long, T)]</pre> <p>Retrieves (<i>gets</i>) metadata of one or more batches</p> <p>Used when...FIXME</p>

getLatest	<code>getLatest(): Option[(Long, T)]</code>  Retrieves the latest-committed metadata (if available) Used when...FIXME
purge	<code>purge(thresholdBatchId: Long): Unit</code>  Used when...FIXME

Note	<p><a href="#">HDFSMetadataLog</a> is the only direct implementation of the <a href="#">MetadataLog Contract</a> in Spark Structured Streaming.</p>
------	---

# HDFSMetadataLog — Hadoop DFS-based Metadata Storage

`HDFSMetadataLog` is a concrete [metadata storage](#) (of type `T`) that uses Hadoop DFS for fault-tolerance and reliability.

`HDFSMetadataLog` uses the given [path](#) as the **metadata directory** with metadata logs. The path is immediately converted to a Hadoop [Path](#) for file management.

`HDFSMetadataLog` uses [Json4s](#) with the [Jackson](#) binding for metadata [serialization](#) and [deserialization](#) (to and from JSON format).

`HDFSMetadataLog` is further customized by the [extensions](#).

Table 1. HDFSMetadataLogs (Direct Extensions Only)

HDFSMetadataLog	Description
<code>Anonymous</code>	<code>HDFSMetadataLog</code> of <a href="#">KafkaSourceOffsets</a> for <a href="#">KafkaSource</a>
<code>Anonymous</code>	<code>HDFSMetadataLog</code> of <a href="#">LongOffsets</a> for <a href="#">RateStreamMicroBatchReader</a>
<code>CommitLog</code>	Offset commit log of <a href="#">streaming query execution engines</a>
<code>CompactibleFileStreamLog</code>	Compactible metadata logs (that compact logs at regular interval)
<code>KafkaSourceInitialOffsetWriter</code>	<code>HDFSMetadataLog</code> of <a href="#">KafkaSourceOffsets</a> for <a href="#">KafkaSource</a>
<code>OffsetSeqLog</code>	Write-Ahead Log ( <a href="#">WAL</a> ) of <a href="#">stream execution engines</a>

## Creating HDFSMetadataLog Instance

`HDFSMetadataLog` takes the following to be created:

- `SparkSession`
- Path of the metadata log directory

While being [created](#) `HDFSMetadataLog` creates the [path](#) unless exists already.

## Serializing Metadata (Writing Metadata in Serialized Format) — `serialize` Method

```
serialize(
  metadata: T,
  out: OutputStream): Unit
```

`serialize` simply writes the log data (serialized using [Json4s](#) (with Jackson binding) library).

Note

`serialize` is used exclusively when `HDFSMetadataLog` is requested to [write metadata of a streaming batch to a file \(metadata log\)](#) (when [storing metadata of a streaming batch](#)).

## Deserializing Metadata (Reading Metadata from Serialized Format) — `deserialize` Method

```
deserialize(in: InputStream): T
```

`deserialize` deserializes a metadata (of type `T`) from a given `InputStream`.

Note

`deserialize` is used exclusively when `HDFSMetadataLog` is requested to [retrieve metadata of a batch](#).

## Retrieving Metadata Of Streaming Batch — `get` Method

```
get(batchId: Long): Option[T]
```

Note

`get` is part of the [MetadataLog Contract](#) to get metadata of a batch.

`get` ...FIXME

## Retrieving Metadata of Range of Batches — `get` Method

```
get(
  startId: Option[Long],
  endId: Option[Long]): Array[(Long, T)]
```

Note

`get` is part of the [MetadataLog Contract](#) to get metadata of range of batches.

```
get ...FIXME
```

## Persisting Metadata of Streaming Micro-Batch — `add` Method

```
add(  
    batchId: Long,  
    metadata: T): Boolean
```

Note	<code>add</code> is part of the <a href="#">MetadataLog Contract</a> to persist metadata of a streaming batch.
------	--

`add` return `true` when the metadata of the streaming batch was not available and persisted successfully. Otherwise, `add` returns `false`.

Internally, `add` looks up metadata of the given streaming batch (`batchId`) and returns `false` when found.

Otherwise, when not found, `add` creates a metadata log file for the given `batchId` and writes metadata to the file. `add` returns `true` if successful.

## Latest Committed Batch Id with Metadata (When Available) — `getLatest` Method

```
getLatest(): Option[(Long, T)]
```

Note	<code>getLatest</code> is a part of <a href="#">MetadataLog Contract</a> to retrieve the recently-committed batch id and the corresponding metadata if available in the metadata storage.
------	---

`getLatest` requests the internal [FileManager](#) for the files in [metadata directory](#) that match [batch file filter](#).

`getLatest` takes the batch ids (the batch files correspond to) and sorts the ids in reverse order.

`getLatest` gives the first batch id with the metadata which could be found in the metadata storage.

Note	It is possible that the batch id could be in the metadata storage, but not available for retrieval.
------	---

## Removing Expired Metadata (Purging) — `purge` Method

```
purge(thresholdBatchId: Long): Unit
```

**Note**

`purge` is part of the [MetadataLog Contract](#) to...FIXME.

`purge` ...FIXME

## Creating Batch Metadata File — `batchIdToPath` Method

```
batchIdToPath(batchId: Long): Path
```

`batchIdToPath` simply creates a Hadoop [Path](#) for the file called by the specified `batchId` under the [metadata directory](#).

**Note**

`batchIdToPath` is used when:

- `CompatibleFileStreamLog` is requested to [compact](#) and [allFiles](#)
- `HDFSMetadataLog` is requested to [add](#), [get](#), [purge](#), and [purgeAfter](#)

## `isBatchFile` Method

```
isBatchFile(path: Path): Boolean
```

`isBatchFile` ...FIXME

**Note**

`isBatchFile` is used exclusively when `HDFSMetadataLog` is requested for the [PathFilter of batch files](#).

## `pathToBatchId` Method

```
pathToBatchId(path: Path): Long
```

`pathToBatchId` ...FIXME

**Note**

`pathToBatchId` is used when:

- `CompatibleFileStreamLog` is requested for the [compact interval](#)
- `HDFSMetadataLog` is requested to [isBatchFile](#), [get metadata of a range of batches](#), [getLatest](#), [getOrderedBatchFiles](#), [purge](#), and [purgeAfter](#)

## verifyBatchIds Object Method

```
verifyBatchIds(  
    batchIds: Seq[Long],  
    startId: Option[Long],  
    endId: Option[Long]): Unit
```

verifyBatchIds ...FIXME

Note

- `verifyBatchIds` is used when:
- `FileStreamSourceLog` is requested to `get`
  - `HDFSMetadataLog` is requested to `get`

## Retrieving Version (From Text Line) — parseVersion Internal Method

```
parseVersion(  
    text: String,  
    maxSupportedVersion: Int): Int
```

parseVersion ...FIXME

Note

- `parseVersion` is used when:
- `KafkaSourceInitialOffsetWriter` is requested to `deserialize metadata`
  - `KafkaSource` is requested for the `initial partition offsets`
  - `CommitLog` is requested to `deserialize metadata`
  - `CompactibleFileStreamLog` is requested to `deserialize metadata`
  - `OffsetSeqLog` is requested to `deserialize metadata`
  - `RateStreamMicroBatchReader` is requested to `deserialize metadata`

## purgeAfter Method

```
purgeAfter(thresholdBatchId: Long): Unit
```

purgeAfter ...FIXME

Note

`purgeAfter` seems to be used exclusively in tests.

## Writing Batch Metadata to File (Metadata Log)

### — `writeBatchToFile` Internal Method

```
writeBatchToFile(  
    metadata: T,  
    path: Path): Unit
```

`writeBatchToFile` requests the [CheckpointFileManager](#) to [createAtomic](#) (for the specified `path` and the `overwriteIfPossible` flag disabled).

`writeBatchToFile` then [serializes the metadata](#) (to the `CancellableFSDataOutputStream` output stream) and closes the stream.

In case of an exception, `writeBatchToFile` simply requests the `CancellableFSDataOutputStream` output stream to `cancel` (so that the output file is not generated) and re-throws the exception.

Note	<code>writeBatchToFile</code> is used exclusively when <code>HDFSMetadataLog</code> is requested to <a href="#">store (persist) metadata of a streaming batch</a> .
------	---

## Retrieving Ordered Batch Metadata Files

### — `getOrderedBatchFiles` Method

```
getOrderedBatchFiles(): Array[FileStatus]
```

`getOrderedBatchFiles` ...FIXME

Note	<code>getOrderedBatchFiles</code> does not seem to be used at all.
------	--

## Internal Properties

Name	Description
batchFilesFilter	<p>Hadoop's <a href="#">PathFilter</a> of batch files (with names being long numbers)</p> <p>Used when:</p> <ul style="list-style-type: none"><li>• <code>CompactibleFileStreamLog</code> is requested for the <code>compactInterval</code></li><li>• <code>HDFSMetadataLog</code> is requested to <a href="#">get batch metadata</a>, <a href="#">getLatest</a>, <a href="#">getOrderedBatchFiles</a>, <a href="#">purge</a>, and <a href="#">purgeAfter</a></li></ul>
fileManager	<p><a href="#">CheckpointFileManager</a></p> <p>Used when...FIXME</p>

# CommitLog — HDFSMetadataLog for Offset Commit Log

`CommitLog` is an [HDFSMetadataLog](#) with [CommitMetadata](#) metadata.

`CommitLog` is [created](#) exclusively for the [offset commit log](#) of [StreamExecution](#).

`CommitLog` uses `CommitMetadata` for the metadata with **nextBatchWatermarkMs** attribute (of type `Long` and the default `0`).

`CommitLog` [writes](#) commit metadata to files with names that are offsets.

```
$ ls -tr [checkpoint-directory]/commits
0 1 2 3 4 5 6 7 8 9

$ cat [checkpoint-directory]/commits/8
v1
{"nextBatchWatermarkMs": 0}
```

`CommitLog` uses **1** for the version.

`CommitLog` (like the parent [HDFSMetadataLog](#)) takes the following to be created:

- `SparkSession`
- Path of the metadata log directory

## Serializing Metadata (Writing Metadata to Persistent Storage) — `serialize` Method

```
serialize(
    metadata: CommitMetadata,
    out: OutputStream): Unit
```

Note	<code>serialize</code> is part of <a href="#">HDFSMetadataLog Contract</a> to write a metadata in serialized format.
------	--

`serialize` writes out the [version](#) prefixed with `v` on a single line (e.g. `v1`) followed by the given `CommitMetadata` in JSON format.

## Deserializing Metadata — `deserialize` Method

```
deserialize(in: InputStream): CommitMetadata
```

**Note**

`deserialize` is part of [HDFSMetadataLog Contract](#) to deserialize a metadata (from an `InputStream` ).

`deserialize` simply reads (`deserializes`) two lines from the given `InputStream` for `version` and the `nextBatchWatermarkMs` attribute.

## **add Method**

```
add(batchId: Long): Unit
```

`add` ...FIXME

**Note**

`add` is used when...FIXME

## **add Method**

```
add(batchId: Long, metadata: String): Boolean
```

**Note**

`add` is part of [MetadataLog Contract](#) to...FIXME.

`add` ...FIXME

# CommitMetadata

CommitMetadata is...FIXME

# OffsetSeqLog — Hadoop DFS-based Metadata Storage of OffsetSeqs

`OffsetSeqLog` is a [Hadoop DFS-based metadata storage](#) for `OffsetSeq` metadata.

`OffsetSeqLog` uses `OffsetSeq` for metadata which holds an ordered collection of offsets and optional metadata (as `OffsetSeqMetadata` for event-time watermark).

`OffsetSeqLog` is [created](#) exclusively for the [write-ahead log \(WAL\)](#) of offsets of [stream execution engines](#) (i.e. `ContinuousExecution` and `MicroBatchExecution`).

`OffsetSeqLog` uses `1` for the version when [serializing](#) and [deserializing](#) metadata.

## Creating OffsetSeqLog Instance

`OffsetSeqLog` (like the parent [HDFSMetadataLog](#)) takes the following to be created:

- `SparkSession`
- Path of the metadata log directory

## Serializing Metadata (Writing Metadata in Serialized Format) — `serialize` Method

```
serialize(  
    offsetSeq: OffsetSeq,  
    out: OutputStream): Unit
```

Note	<code>serialize</code> is part of <a href="#">HDFSMetadataLog Contract</a> to serialize metadata (write metadata in serialized format).
------	---

`serialize` firstly writes out the `version` prefixed with `v` on a single line (e.g. `v1`) followed by the [optional metadata](#) in JSON format.

`serialize` then writes out the `offsets` in JSON format, one per line.

Note	No offsets to write in <code>offsetSeq</code> for a streaming source is marked as <code>-</code> (a dash) in the log.
------	---

```
$ ls -tr [checkpoint-directory]/offsets
0 1 2 3 4 5 6

$ cat [checkpoint-directory]/offsets/6
v1
{"batchWatermarkMs":0,"batchTimestampMs":1502872590006,"conf":{"spark.sql.shuffle.partitions":"200","spark.sql.streaming.stateStore.providerClass":"org.apache.spark.sql.execution.streaming.state.HDFSBackedStateStoreProvider"}}
51
```

## Deserializing Metadata (Reading OffsetSeq from Serialized Format) — `deserialize` Method

`deserialize(in: InputStream): OffsetSeq`

Note	<code>deserialize</code> is part of <a href="#">HDFSMetadataLog Contract</a> to deserialize metadata (read metadata from serialized format).
------	--

`deserialize` firstly parses the [version](#) on the first line.

`deserialize` reads the optional metadata (with an empty line for metadata not available).

`deserialize` creates a [SerializedOffset](#) for every line left.

In the end, `deserialize` creates a [OffsetSeq](#) for the optional metadata and the [SerializedOffsets](#).

When there are no lines in the [InputStream](#), `deserialize` throws an [IllegalStateException](#):

Incomplete log file

# OffsetSeq

`OffsetSeq` is the metadata managed by [Hadoop DFS-based metadata storage](#).

`OffsetSeq` is [created](#) (possibly using the `fill` factory methods) when:

- `OffsetSeqLog` is requested to [deserialize metadata](#) (retrieve metadata from a persistent storage)
- `StreamProgress` is requested to [convert itself to OffsetSeq](#) (most importantly when `MicroBatchExecution` stream execution engine is requested to [construct the next streaming micro-batch](#) to [commit available offsets for a batch to the write-ahead log](#))
- `ContinuousExecution` stream execution engine is requested to [get start offsets](#) and [addOffset](#)

## Creating OffsetSeq Instance

`OffsetSeq` takes the following when created:

- Collection of optional [Offsets](#) (with `None` for [streaming sources with no new data available](#))
- Optional [OffsetSeqMetadata](#) (default: `None`)

## Converting to StreamProgress — `toStreamProgress` Method

```
toStreamProgress(  
    sources: Seq[BaseStreamingSource]): StreamProgress
```

`toStreamProgress` creates a new [StreamProgress](#) and adds the [streaming sources](#) for which there are new [offsets](#) available.

Note	<a href="#">Offsets</a> is a collection with <i>holes</i> (empty elements) for streaming sources with no new data available.
------	--

`toStreamProgress` throws an `AssertionError` if the number of the input `sources` does not match the [offsets](#):

```
There are [[offsets.size]] sources in the checkpoint offsets and now there are [[sources.size]] sources requested by the query. Cannot continue.
```

**Note**

- `toStreamProgress` is used when:
- `MicroBatchExecution` is requested to [populate start offsets from offsets and commits checkpoints](#) and [construct \(or skip\) the next streaming micro-batch](#)
  - `ContinuousExecution` is requested for [start offsets](#)

## Textual Representation — `toString` Method

```
toString: String
```

**Note**

`toString` is part of the [java.lang.Object](#) contract for the string representation of the object.

`toString` simply converts the [Offsets](#) to JSON (if an offset is available) or `-` (a dash if an offset is not available for a streaming source at that position).

## Creating OffsetSeq Instance — `fill` Factory Methods

```
fill(
  offsets: Offset*): OffsetSeq (1)
fill(
  metadata: Option[String],
  offsets: Offset*): OffsetSeq
```

1. Uses no metadata (`None`)

`fill` simply creates an [OffsetSeq](#) for the given variable sequence of [Offsets](#) and the optional [OffsetSeqMetadata](#) (in JSON format).

**Note**

- `fill` is used when:
- `OffsetSeqLog` is requested to [deserialize metadata](#)
  - `ContinuousExecution` stream execution engine is requested to [get start offsets](#) and [addOffset](#)

# CompactibleFileStreamLog Contract — Compactible Metadata Logs

`CompactibleFileStreamLog` is the [extension](#) of the `HDFSMetadataLog` contract for [compatible metadata logs](#) that [compactLogs](#) every [compact interval](#).

`CompactibleFileStreamLog` uses `spark.sql.streaming.minBatchesToRetain` configuration property (default: `100`) for [deleteExpiredLog](#).

`CompactibleFileStreamLog` uses `.compact` suffix for `batchIdToPath`, `getBatchIdFromFileName`, and the `compactInterval`.

Table 1. CompactibleFileStreamLog Contract (Abstract Methods Only)

Method	Description
<code>compactLogs</code>	<code>compactLogs(logs: Seq[T]): Seq[T]</code>  Used when <code>CompactibleFileStreamLog</code> is requested to <a href="#">compact</a> and <a href="#">allFiles</a>
<code>defaultCompactInterval</code>	<code>defaultCompactInterval: Int</code>  Default <a href="#">compaction interval</a>  Used exclusively when <code>CompactibleFileStreamLog</code> is requested for the <code>compactInterval</code>
<code>fileCleanupDelayMs</code>	<code>fileCleanupDelayMs: Long</code>  Used exclusively when <code>CompactibleFileStreamLog</code> is requested to <a href="#">deleteExpiredLog</a>
<code>isDeletingExpiredLog</code>	<code>isDeletingExpiredLog: Boolean</code>  Used exclusively when <code>CompactibleFileStreamLog</code> is requested to <a href="#">store (add)</a> metadata of a streaming batch

Table 2. CompactibleFileStreamLogs

CompactibleFileStreamLog	Description
<a href="#">FileStreamSinkLog</a>	
<a href="#">FileStreamSourceLog</a>	CompactibleFileStreamLog (of <code>FileEntry</code> metadata) of <a href="#">FileStreamSource</a>

## Creating CompactibleFileStreamLog Instance

`CompactibleFileStreamLog` takes the following to be created:

- Metadata version
- `SparkSession`
- Path of the metadata log directory

Note	<code>CompactibleFileStreamLog</code> is a Scala abstract class and cannot be created directly. It is created indirectly for the <a href="#">concrete CompactibleFileStreamLogs</a> .
------	---

### batchIdToPath Method

```
batchIdToPath(batchId: Long): Path
```

Note	<code>batchIdToPath</code> is part of the <a href="#">HDFSMetadataLog Contract</a> to...FIXME.
------	--

`batchIdToPath` ...FIXME

### pathToBatchId Method

```
pathToBatchId(path: Path): Long
```

Note	<code>pathToBatchId</code> is part of the <a href="#">HDFSMetadataLog Contract</a> to...FIXME.
------	--

`pathToBatchId` ...FIXME

### isBatchFile Method

```
isBatchFile(path: Path): Boolean
```

**Note**

`isBatchFile` is part of the [HDFSMetadataLog Contract](#) to...FIXME.

`isBatchFile` ...FIXME

## Serializing Metadata (Writing Metadata in Serialized Format) — `serialize` Method

```
serialize(  
    logData: Array[T],  
    out: OutputStream): Unit
```

**Note**

`serialize` is part of the [HDFSMetadataLog Contract](#) to serialize metadata (write metadata in serialized format).

`serialize` firstly writes the version header (`v` and the `metadataLogVersion`) out to the given output stream (in `UTF_8` ).

`serialize` then writes the log data (serialized using [Json4s \(with Jackson binding\)](#) library). Entries are separated by new lines.

## Deserializing Metadata — `deserialize` Method

```
deserialize(in: InputStream): Array[T]
```

**Note**

`deserialize` is part of the [HDFSMetadataLog Contract](#) to...FIXME.

`deserialize` ...FIXME

## Storing Metadata Of Streaming Batch — `add` Method

```
add(  
    batchId: Long,  
    logs: Array[T]): Boolean
```

**Note**

`add` is part of the [HDFSMetadataLog Contract](#) to store metadata for a batch.

`add` ...FIXME

## allFiles Method

```
allFiles(): Array[T]
```

`allFiles` ...FIXME

#### Note

- `allFiles` is used when:
- `FileStreamSource` is created
  - `MetadataLogFileIndex` is created

## compact Internal Method

```
compact(
  batchId: Long,
  logs: Array[T]): Boolean
```

`compact` [getValidBatchesBeforeCompactionBatch](#) (with the streaming batch and the [compact interval](#)).

`compact` ...FIXME

In the end, `compact` [compactLogs](#) and requests the parent `HDFSMetadataLog` to [persist](#) metadata of a streaming batch (to a metadata log file).

#### Note

`compact` is used exclusively when `CompactibleFileStreamLog` is requested to [persist metadata of a streaming batch](#).

## getValidBatchesBeforeCompactionBatch Object Method

```
getValidBatchesBeforeCompactionBatch(
  compactionBatchId: Long,
  compactInterval: Int): Seq[Long]
```

`getValidBatchesBeforeCompactionBatch` ...FIXME

#### Note

`getValidBatchesBeforeCompactionBatch` is used exclusively when `CompactibleFileStreamLog` is requested to [compact](#).

## isCompactionBatch Object Method

```
isCompactionBatch(batchId: Long, compactInterval: Int): Boolean
```

`isCompactionBatch` ...FIXME

#### Note

- `isCompactionBatch` is used when:
  - `CompactibleFileStreamLog` is requested to `batchIdToPath`, `store the metadata of a batch`, `deleteExpiredLog`, and `getValidBatchesBeforeCompactionBatch`
  - `FileStreamSourceLog` is requested to `store the metadata of a batch` and get

## getBatchIdFromFile Name Object Method

```
getBatchIdFromFile Name(fileName: String): Long
```

`getBatchIdFromFile Name` simply removes the `.compact` suffix from the given `fileName` and converts the remaining part to a number.

#### Note

- `getBatchIdFromFile Name` is used when `CompactibleFileStreamLog` is requested to `pathToBatchId`, `isBatchFile`, and `deleteExpiredLog`.

## deleteExpiredLog Internal Method

```
deleteExpiredLog(
  currentBatchId: Long): Unit
```

`deleteExpiredLog` does nothing and simply returns when the current batch ID incremented (`currentBatchId + 1`) is below the `compact interval` plus the `minBatchesToRetain`.

`deleteExpiredLog` ...FIXME

#### Note

- `deleteExpiredLog` is used exclusively when `CompactibleFileStreamLog` is requested to `store metadata of a streaming batch`.

## Internal Properties

Name	Description
<code>compactInterval</code>	<b>Compact interval</b>



# FileStreamSourceLog

`FileStreamSourceLog` is a concrete `CompactibleFileStreamLog` (of `FileEntry` metadata) of `FileStreamSource`.

`FileStreamSourceLog` uses a fixed-size `cache` of metadata of compaction batches.

`FileStreamSourceLog` uses `spark.sql.streaming.fileSource.log.compactInterval` configuration property (default: `10`) for the `default` compaction interval.

`FileStreamSourceLog` uses `spark.sql.streaming.fileSource.log.cleanupDelay` configuration property (default: `10` minutes) for the `fileCleanupDelayMs`.

`FileStreamSourceLog` uses `spark.sql.streaming.fileSource.log.deletion` configuration property (default: `true`) for the `isDeletingExpiredLog`.

## Creating FileStreamSourceLog Instance

`FileStreamSourceLog` (like the parent `CompactibleFileStreamLog`) takes the following to be created:

- Metadata version
- `SparkSession`
- Path of the metadata log directory

## Storing (Adding) Metadata of Streaming Batch — `add` Method

```
add(
  batchId: Long,
  logs: Array[FileEntry]): Boolean
```

Note

`add` is part of the `MetadataLog Contract` to store (`add`) metadata of a streaming batch.

`add` requests the parent `CompactibleFileStreamLog` to `store` metadata (possibly `compacting` logs if the batch is `compaction`).

If so (and this is a compaction batch), `add` adds the batch and the logs to `fileEntryCache` internal registry (and possibly removing the eldest entry if the size is above the `cacheSize`).

## get Method

```
get(
  startId: Option[Long],
  endId: Option[Long]): Array[(Long, Array[FileEntry])]
```

Note

`get` is part of the [MetadataLog Contract](#) to...FIXME.

`get` ...FIXME

## Internal Properties

Name	Description
<code>cacheSize</code>	<p>Size of the <a href="#">fileEntryCache</a> that is exactly the <a href="#">compact interval</a></p> <p>Used when the <a href="#">fileEntryCache</a> is requested to add a new entry in <a href="#">add</a> and <a href="#">get</a> a compaction batch</p>
<code>fileEntryCache</code>	<p>Metadata of a streaming batch ( <a href="#">FileEntry</a> ) per batch ID ( <a href="#">LinkedHashMap[Long, Array[FileEntry]]</a> ) of size configured using the <a href="#">cacheSize</a></p> <ul style="list-style-type: none"> <li>• New entry added for a <a href="#">compaction batch</a> when <a href="#">storing (adding) metadata of a streaming batch</a></li> </ul> <p>Used when <a href="#">get</a> (for a <a href="#">compaction batch</a>)</p>

# OffsetSeqMetadata — Metadata of Streaming Batch

`OffsetSeqMetadata` holds the metadata for the current streaming batch:

- Event-time watermark threshold
- Batch timestamp (in millis)
- **Streaming configuration** with `spark.sql.shuffle.partitions` and `spark.sql.streaming.stateStore.providerClass` Spark properties

Note	<code>OffsetSeqMetadata</code> is used mainly when <code>IncrementalExecution</code> is created.
------	--

`OffsetSeqMetadata` considers some configuration properties as **relevantSQLConfs**:

- `SHUFFLE_PARTITIONS`
- `STATE_STORE_PROVIDER_CLASS`
- `STREAMING_MULTIPLE_WATERMARK_POLICY`
- `FLATMAPGROUPSWITHSTATE_STATE_FORMAT_VERSION`
- `STREAMING_AGGREGATION_STATE_FORMAT_VERSION`

`relevantSQLConfs` are used when `OffsetSeqMetadata` is created and is requested to `setSessionConf`.

## Creating OffsetSeqMetadata — apply Factory Method

```
apply(  
    batchWatermarkMs: Long,  
    batchTimestampMs: Long,  
    sessionConf: RuntimeConfig): OffsetSeqMetadata
```

`apply ...FIXME`

Note	<code>apply</code> is used when...FIXME
------	---

## setSessionConf Method

```
setSessionConf(metadata: OffsetSeqMetadata, sessionConf: RuntimeConfig): Unit
```

`setSessionConf ...FIXME`

<b>Note</b>	<code>setSessionConf</code> is used when...FIXME
-------------	--

# CheckpointFileManager Contract

`CheckpointFileManager` is the abstraction of [checkpoint managers](#) that manage checkpoint files (metadata of streaming batches) on Hadoop DFS-compatible file systems.

`CheckpointFileManager` is created per [spark.sql.streaming.checkpointFileManagerClass](#) configuration property if defined before reverting to the available [checkpoint managers](#).

`CheckpointFileManager` is used exclusively by [HDFSMetadataLog](#), [StreamMetadata](#) and [HDFSBackedStateStoreProvider](#).

Table 1. CheckpointFileManager Contract

Method	Description
<code>createAtomic</code>	<pre>createAtomic(     path: Path,     overwriteIfPossible: Boolean): CancellableFSDataOutputSteam</pre> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>HDFSMetadataLog</code> is requested to <a href="#">store metadata for a batch</a> (that <a href="#">writeBatchToFile</a>)</li> <li>• <code>StreamMetadata</code> helper object is requested to <a href="#">persist metadata</a></li> <li>• <code>HDFSBackedStateStore</code> is requested for the <a href="#">deltaFileStream</a></li> <li>• <code>HDFSBackedStateStoreProvider</code> is requested to <a href="#">writeSnapshotFile</a></li> </ul>
<code>delete</code>	<pre>delete(path: Path): Unit</pre> <p>Deletes the given path recursively (if exists)</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>RenameBasedFSDataOutputStream</code> is requested to <a href="#">cancel</a></li> <li>• <code>CompactibleFileStreamLog</code> is requested to <a href="#">store metadata for a batch</a> (that <a href="#">deleteExpiredLog</a>)</li> <li>• <code>HDFSMetadataLog</code> is requested to <a href="#">remove expired metadata</a> and <a href="#">purgeAfter</a></li> <li>• <code>HDFSBackedStateStoreProvider</code> is requested to <a href="#">do maintenance</a> (that <a href="#">cleans up</a>)</li> </ul>

	<code>exists(path: Path): Boolean</code>
<code>exists</code>	Used when <code>HDFSMetadataLog</code> is created (to create the metadata directory) and requested for metadata of a batch
<code>isLocal</code>	<code>isLocal: Boolean</code>  Does not seem to be used.
<code>list</code>	<code>list(     path: Path): Array[FileStatus] (1) list(     path: Path,     filter: PathFilter): Array[FileStatus]</code>  1. Uses <code>PathFilter</code> that accepts all files in the path  Lists all files in the given path  Used when: <ul style="list-style-type: none"><li>• <code>HDFSBackedStateStoreProvider</code> is requested for all delta and snapshot files</li><li>• <code>CompactibleFileStreamLog</code> is requested for the compact interval and to <code>deleteExpiredLog</code></li><li>• <code>HDFSMetadataLog</code> is requested for metadata of one or more batches, the latest committed batch, ordered batch metadata files, to remove expired metadata and <code>purgeAfter</code></li></ul>
<code>mkdirs</code>	<code>mkdirs(path: Path): Unit</code>  Used when: <ul style="list-style-type: none"><li>• <code>HDFSMetadataLog</code> is created</li><li>• <code>HDFSBackedStateStoreProvider</code> is requested to initialize</li></ul>
<code>open</code>	<code>open(path: Path): FSDataInputStream</code>  Opens a file (by the given path) for reading  Used when: <ul style="list-style-type: none"><li>• <code>HDFSMetadataLog</code> is requested for metadata of a batch</li></ul>

- `HDFSBackedStateStoreProvider` is requested to [retrieve the state store for a specified version](#) (that `updateFromDeltaFile`), and `readSnapshotFile`

Table 2. CheckpointFileManagers

CheckpointFileManager	Description
<code>FileContextBasedCheckpointFileManager</code>	Default <code>CheckpointFileManager</code> that uses Hadoop's <code>FileContext</code> API for managing checkpoint files (unless <code>spark.sql.streaming.checkpointFileManagerC</code> configuration property is used)
<code>FileSystemBasedCheckpointFileManager</code>	Basic <code>CheckpointFileManager</code> that uses Hadoop's <code>FileSystem</code> API for managing checkpoint files (that <a href="#">assumes</a> that the implementation of <code>FileSystem.rename()</code> is atomic or the correctness and fault-tolerance Structured Streaming is not guaranteed)

## Creating CheckpointFileManager Instance — `create` Object Method

```
create(
  path: Path,
  hadoopConf: Configuration): CheckpointFileManager
```

`create` finds `spark.sql.streaming.checkpointFileManagerClass` configuration property in the `hadoopConf` configuration.

If found, `create` simply instantiates whatever `CheckpointFileManager` implementation is defined.

If not found, `create` creates a `FileContextBasedCheckpointFileManager`.

In case of `UnsupportedFileSystemException`, `create` prints out the following WARN message to the logs and creates (*falls back on*) a `FileSystemBasedCheckpointFileManager`.

```
Could not use FileContext API for managing Structured Streaming checkpoint files at [path]. Using FileSystem API instead for managing log files. If the implementation of FileSystem.rename() is not atomic, then the correctness and fault-tolerance of your Structured Streaming is not guaranteed.
```

**Note**

`create` is used when:

- `HDFSMetadataLog` is [created](#)
- `StreamMetadata` helper object is requested to [write metadata to a file](#) (when `StreamExecution` is [created](#))
- `HDFSBackedStateStoreProvider` is requested for the [CheckpointFileManager](#)

# FileContextBasedCheckpointFileManager

FileContextBasedCheckpointFileManager is...FIXME

# FileSystemBasedCheckpointFileManager — CheckpointFileManager on Hadoop's FileSystem API

`FileSystemBasedCheckpointFileManager` is a [CheckpointFileManager](#) that uses Hadoop's [FileSystem](#) API for managing checkpoint files:

- `list` uses [FileSystem.listStatus](#)
- `mkdirs` uses [FileSystem.mkdirs](#)
- `createTempFile` uses [FileSystem.create](#) (with overwrite enabled)
- `createAtomic` uses [RenameBasedFSDataOutputStream](#)
- `open` uses [FileSystem.open](#)
- `exists` uses [FileSystem.getFileStatus](#)
- `renameTempFile` uses [FileSystem.rename](#)
- `delete` uses [FileSystem.delete](#) (with recursive enabled)
- `isLocal` is `true` for the [FileSystem](#) being [LocalFileSystem](#) or [RawLocalFileSystem](#)

`FileSystemBasedCheckpointFileManager` is [created](#) exclusively when [CheckpointFileManager](#) helper object is requested for a [CheckpointFileManager](#) (for [HDFSMetadataLog](#), [StreamMetadata](#) and [HDFSBackedStateStoreProvider](#)).

`FileSystemBasedCheckpointFileManager` is a [RenameHelperMethods](#) for [atomicity](#) by "write-to-temp-file-and-rename".

## Creating FileSystemBasedCheckpointFileManager Instance

`FileSystemBasedCheckpointFileManager` takes the following to be created:

- Checkpoint directory (Hadoop's [Path](#))
- Configuration (Hadoop's [Configuration](#))

`FileSystemBasedCheckpointFileManager` initializes the [internal properties](#).

## Internal Properties

Name	Description
fs	Hadoop's <a href="#">FileSystem</a> of the <a href="#">checkpoint</a> directory

# Offset—Read Position of Streaming Query

`offset` is the [base](#) of [stream positions](#) that represent progress of a streaming query in [json](#) format.

Table 1. Offset Contract (Abstract Methods Only)

Method	Description
<code>json</code>	<pre>String json()</pre> <p>Converts the offset to JSON format (JSON-encoded offset)</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>MicroBatchExecution</code> stream execution engine is requested to <a href="#">construct the next streaming micro-batch</a> and <a href="#">run a streaming micro-batch</a> (with <a href="#">MicroBatchReader</a> sources)</li> <li>• <code>offsetSeq</code> is requested for the <a href="#">textual representation</a></li> <li>• <code>offsetSeqLog</code> is requested to <a href="#">serialize metadata</a> (<a href="#">write metadata in serialized format</a>)</li> <li>• <code>ProgressReporter</code> is requested to <a href="#">record trigger offsets</a></li> <li>• <code>ContinuousExecution</code> stream execution engine is requested to <a href="#">run a streaming query in continuous mode</a> and <a href="#">commit an epoch</a></li> </ul>

Table 2. Offsets

Offset	Description
ContinuousMemoryStreamOffset	
FileStreamSourceOffset	
KafkaSourceOffset	
LongOffset	
RateStreamOffset	
SerializedOffset	JSON-encoded offset that is used when loading an offset from an external storage, e.g. from <a href="#">checkpoint</a> after restart
TextSocketOffset	

# StreamProgress — Collection of Offsets per Streaming Source

`StreamProgress` is a collection of [Offsets](#) per streaming source.

`StreamProgress` is [created](#) when:

- `StreamExecution` is [created](#) (and creates [committed](#) and [available](#) offsets)
- `OffsetSeq` is requested to [convert to StreamProgress](#)

`StreamProgress` is an extension of Scala's `scala.collection.immutable.Map` with [streaming sources](#) as keys and their [Offsets](#) as values.

## Creating StreamProgress Instance

`StreamProgress` takes the following to be created:

- Optional collection of [offsets](#) per [streaming source](#) (`Map[BaseStreamingSource, Offset]`) (default: empty)

## Looking Up Offset by Streaming Source — `get` Method

```
get(key: BaseStreamingSource): Option[Offset]
```

Note	<code>get</code> is part of the Scala's <code>scala.collection.MapLike</code> to...FIXME.
------	---

`get` simply looks up an [Offsets](#) for the given [BaseStreamingSource](#) in the [baseMap](#).

## `++` Method

```
++(  
  updates: GenTraversableOnce[(BaseStreamingSource, Offset)]): StreamProgress
```

`++` simply creates a new [StreamProgress](#) with the [baseMap](#) and the given [updates](#).

Note	<code>++</code> is used exclusively when <code>OffsetSeq</code> is requested to <a href="#">convert to StreamProgress</a> .
------	---

## Converting to OffsetSeq — `toOffsetSeq` Method

```
toOffsetSeq(  
    sources: Seq[BaseStreamingSource],  
    metadata: OffsetSeqMetadata): OffsetSeq
```

`toOffsetSeq` creates a [OffsetSeq](#) with offsets that are [looked up](#) for every [BaseStreamingSource](#).

Note	<p><code>toOffsetSeq</code> is used when:</p> <ul style="list-style-type: none"><li>• <code>MicroBatchExecution</code> stream execution engine is requested to <a href="#">construct the next streaming micro-batch</a> (to <a href="#">commit available offsets</a> for a batch to the write-ahead log)</li><li>• <code>StreamExecution</code> is requested to <a href="#">run stream processing</a> (that <a href="#">failed with a Throwable</a>)</li></ul>
------	--

# Micro-Batch Stream Processing (Structured Streaming V1)

**Micro-Batch Stream Processing** is a stream processing model in Spark Structured Streaming that is used for streaming queries with [Trigger.Once](#) and [Trigger.ProcessingTime](#) triggers.

Micro-batch stream processing uses [MicroBatchExecution](#) stream execution engine.

Micro-batch stream processing supports [MicroBatchReadSupport](#) data sources.

Micro-batch stream processing is often referred to as **Structured Streaming V1**.

```
import org.apache.spark.sql.streaming.Trigger
import scala.concurrent.duration._

val sq = spark
  .readStream
  .format("rate")
  .load
  .writeStream
  .format("console")
  .option("truncate", false)
  .trigger(Trigger.ProcessingTime(1.minute)) // <-- Uses MicroBatchExecution for execution
  .queryName("rate2console")
  .start

assert(sq.isActive)

scala> sq.explain
== Physical Plan ==
WriteToDataSourceV2 org.apache.spark.sql.execution.streaming.sources.MicroBatchWriter@678e6267
+- *(1) Project [timestamp#54, value#55L]
  +- *(1) ScanV2 rate[timestamp#54, value#55L]

// sq.stop
```

## Execution Phases (Processing Cycle)

Once [MicroBatchExecution](#) stream processing engine is requested to run an activated [streaming query](#), the query execution goes through the following **execution phases** every trigger:

1. [triggerExecution](#)

2. `getOffset` for [Sources](#) or `setOffsetRange` for [MicroBatchReaders](#)
3. `getEndOffset`
4. `walCommit`
5. `getBatch`
6. `queryPlanning`
7. `addBatch`

Execution phases with execution times are available using [StreamingQueryProgress](#) under

```
durationMs .
```

```
scala> :type sq
org.apache.spark.sql.streaming.StreamingQuery
sq.lastProgress.durationMs.get("walCommit")
```

Tip

Enable INFO logging level for [StreamExecution](#) logger to be notified about durations.

```
17/08/11 09:04:17 INFO StreamExecution: Streaming query made progress: {
  "id" : "ec8f8228-90f6-4e1f-8ad2-80222affed63",
  "runId" : "f605c134-cfb0-4378-88c1-159d8a7c232e",
  "name" : "rates-to-console",
  "timestamp" : "2017-08-11T07:04:17.373Z",
  "batchId" : 0,
  "numInputRows" : 0,
  "processedRowsPerSecond" : 0.0,
  "durationMs" : {           // <-- Durations (in millis)
    "addBatch" : 38,
    "getBatch" : 1,
    "getOffset" : 0,
    "queryPlanning" : 1,
    "triggerExecution" : 62,
    "walCommit" : 19
  },
}
```

## Monitoring (using [StreamingQueryListener](#) and Logs)

`MicroBatchExecution` [posts events](#) to announce when a streaming query is started and stopped as well as after every micro-batch. [StreamingQueryListener](#) interface can be used to intercept the events and act accordingly.

After [triggerExecution](#) phase `MicroBatchExecution` is requested to [finish](#) up a streaming batch (trigger) and generate a [StreamingQueryProgress](#) (with execution statistics).

`MicroBatchExecution` prints out the following DEBUG message to the logs:

```
Execution stats: [executionStats]
```

`MicroBatchExecution` posts a `QueryProgressEvent` with the `StreamingQueryProgress` and prints out the following INFO message to the logs:

```
Streaming query made progress: [newProgress]
```

# MicroBatchExecution — Stream Execution Engine of Micro-Batch Stream Processing

`MicroBatchExecution` is the [stream execution engine](#) in [Micro-Batch Stream Processing](#).

`MicroBatchExecution` is [created](#) when `StreamingQueryManager` is requested to [create a streaming query](#) (when `DataStreamWriter` is requested to [start an execution of the streaming query](#)) with the following:

- Any type of [sink](#) but [StreamWriterSupport](#)
- Any type of [trigger](#) but [ContinuousTrigger](#)

```
import org.apache.spark.sql.streaming.Trigger
val query = spark
  .readStream
  .format("rate")
  .load
  .writeStream
  .format("console")           // <-- not a StreamWriterSupport sink
  .option("truncate", false)
  .trigger(Trigger.Once)       // <-- Gives MicroBatchExecution
  .queryName("rate2console")
  .start

// The following gives access to the internals
// And to MicroBatchExecution
import org.apache.spark.sql.execution.streaming.StreamingQueryWrapper
val engine = query.asInstanceOf[StreamingQueryWrapper].streamingQuery
import org.apache.spark.sql.execution.streaming.StreamExecution
assert(engine.isInstanceOf[StreamExecution])

import org.apache.spark.sql.execution.streaming.MicroBatchExecution
val microBatchEngine = engine.asInstanceOf[MicroBatchExecution]
assert(microBatchEngine.trigger == Trigger.Once)
```

Once [created](#), `MicroBatchExecution` (as a [stream execution engine](#)) is requested to [run an activated streaming query](#).

**Tip**

Enable `ALL` logging level for `org.apache.spark.sql.execution.streaming.MicroBatchExecution` to see what happens inside.

Add the following line to `conf/log4j.properties` :

```
log4j.logger.org.apache.spark.sql.execution.streaming.MicroBatchExecution=ALL
```

Refer to [Logging](#).

## Creating MicroBatchExecution Instance

`MicroBatchExecution` takes the following to be created:

- `SparkSession`
- Name of the streaming query
- Path of the checkpoint directory
- Analyzed logical query plan of the streaming query ( `LogicalPlan` )
- [Streaming sink](#)
- [Trigger](#)
- Trigger clock ( `Clock` )
- [Output mode](#)
- Extra options ( `Map[String, String]` )
- `deleteCheckpointOnStop` flag to control whether to delete the checkpoint directory on stop

`MicroBatchExecution` initializes the [internal properties](#).

## MicroBatchExecution and TriggerExecutor

### — triggerExecutor Property

```
triggerExecutor: TriggerExecutor
```

`triggerExecutor` is the [TriggerExecutor](#) of the streaming query that is how micro-batches are executed at regular intervals.

`triggerExecutor` is initialized based on the given `Trigger` (that was used to create the `MicroBatchExecution`):

- `ProcessingTimeExecutor` for `Trigger.ProcessingTime`
- `OneTimeExecutor` for `OneTimeTrigger` (aka `Trigger.Once` trigger)

`triggerExecutor` throws an `IllegalStateException` when the `Trigger` is not one of the built-in implementations.

```
Unknown type of trigger: [trigger]
```

Note	<code>triggerExecutor</code> is used exclusively when <code>StreamExecution</code> is requested to run an activated streaming query (at regular intervals).
------	---

## Running Activated Streaming Query — `runActivatedStream` Method

```
runActivatedStream(  
    sparkSessionForStream: SparkSession): Unit
```

Note	<code>runActivatedStream</code> is part of <code>StreamExecution Contract</code> to run the activated streaming query.
------	--

`runActivatedStream` simply requests the `TriggerExecutor` to execute micro-batches using the `batch runner` (until `MicroBatchExecution` is `terminated` due to a query stop or a failure).

## TriggerExecutor's Batch Runner

The batch runner (of the `TriggerExecutor`) is executed as long as the `MicroBatchExecution` is `active`.

Note	<code>trigger</code> and <code>batch</code> are considered equivalent and used interchangeably.
------	---

The batch runner initializes query progress for the new trigger (aka `startTrigger`).

The batch runner starts `triggerExecution` execution phase that is made up of the following steps:

1. Populating start offsets from `checkpoint` before the first "zero" batch (at every start or restart)
2. Constructing or skipping the next streaming micro-batch

### 3. Running the streaming micro-batch

At the start or restart (`resume`) of a streaming query (when the `current batch ID` is uninitialized and `-1`), the batch runner `populates start offsets from checkpoint` and then prints out the following INFO message to the logs (using the `committedOffsets` internal registry):

```
Stream started from [committedOffsets]
```

The batch runner sets the human-readable description for any Spark job submitted (that streaming sources may submit to get new data) as the `batch description`.

The batch runner `constructs the next streaming micro-batch` (when the `isCurrentBatchConstructed` internal flag is off).

The batch runner `records trigger offsets` (with the `committed` and `available` offsets).

The batch runner updates the `current StreamingQueryStatus` with the `isNewDataAvailable` for `isDataAvailable` property.

With the `isCurrentBatchConstructed` flag enabled (`true`), the batch runner `updates the status message` to one of the following (per `isNewDataAvailable`) and `runs the streaming micro-batch`.

```
Processing new data
```

```
No new data but cleaning up state
```

With the `isCurrentBatchConstructed` flag disabled (`false`), the batch runner simply `updates the status message` to the following:

```
Waiting for data to arrive
```

The batch runner `finalizes query progress for the trigger` (with a flag that indicates whether the current batch had new data).

With the `isCurrentBatchConstructed` flag enabled (`true`), the batch runner increments the `currentBatchId` and turns the `isCurrentBatchConstructed` flag off (`false`).

With the `isCurrentBatchConstructed` flag disabled (`false`), the batch runner simply sleeps (as long as configured using the `spark.sql.streaming.pollingDelay` configuration property).

In the end, the batch runner updates the status message to the following status and returns whether the `MicroBatchExecution` is active or not.

```
Waiting for next trigger
```

## Populating Start Offsets From Checkpoint (Resuming from Checkpoint) — `populateStartOffsets` Internal Method

```
populateStartOffsets(  
    sparkSessionToRunBatches: SparkSession): Unit
```

`populateStartOffsets` requests the Offset Write-Ahead Log for the latest committed batch id with metadata (i.e. `OffsetSeq`).

Note	The batch id could not be available in the write-ahead log when a streaming query started with a new log or no batch was persisted (added) to the log before.
------	---

`populateStartOffsets` branches off based on whether the latest committed batch was available or not.

Note	<code>populateStartOffsets</code> is used exclusively when <code>MicroBatchExecution</code> is requested to run an activated streaming query (before the first "zero" micro-batch).
------	---

## Latest Committed Batch Available

When the latest committed batch id with the metadata was available in the Offset Write-Ahead Log, `populateStartOffsets` (re)initializes the internal state as follows:

- Sets the current batch ID to the latest committed batch ID found
- Turns the `isCurrentBatchConstructed` internal flag on ( `true` )
- Sets the available offsets to the offsets (from the metadata)

When the latest batch ID found is greater than `0`, `populateStartOffsets` requests the Offset Write-Ahead Log for the second latest batch ID with metadata or throws an `IllegalStateException` if not found.

```
batch [latestBatchId - 1] doesn't exist
```

`populateStartOffsets` sets the committed offsets to the second latest committed offsets.

`populateStartOffsets` updates the offset metadata.

Caution	FIXME Describe me
---------	-------------------

`populateStartOffsets` requests the [Offset Commit Log](#) for the latest committed batch id with metadata (i.e. [CommitMetadata](#)).

Caution	FIXME Describe me
---------	-------------------

When the latest committed batch id with metadata was found which is exactly the latest batch ID (found in the [Offset Commit Log](#)), `populateStartOffsets` ...FIXME

When the latest committed batch id with metadata was found, but it is not exactly the second latest batch ID (found in the [Offset Commit Log](#)), `populateStartOffsets` prints out the following WARN message to the logs:

```
Batch completion log latest batch id is
[latestCommittedBatchId], which is not trailing batchid
[latestBatchId] by one
```

When no commit log present in the [Offset Commit Log](#), `populateStartOffsets` prints out the following INFO message to the logs:

```
no commit log present
```

In the end, `populateStartOffsets` prints out the following DEBUG message to the logs:

```
Resuming at batch [currentBatchId] with committed offsets
[committedOffsets] and available offsets [availableOffsets]
```

## No Latest Committed Batch

When the latest committed batch id with the metadata could not be found in the [Offset Write-Ahead Log](#), it is assumed that the streaming query is started for the very first time (or the [checkpoint location](#) has changed).

`populateStartOffsets` prints out the following INFO message to the logs:

```
Starting new streaming query.
```

`populateStartOffsets` sets the `current batch ID` to `0` and creates a new `WatermarkTracker`.

## Constructing Or Skipping Next Streaming Micro-Batch — `constructNextBatch` Internal Method

```
constructNextBatch(  
    noDataBatchesEnabled: Boolean): Boolean
```

Note	<code>constructNextBatch</code> will only be executed when the <code>isCurrentBatchConstructed</code> internal flag is enabled ( <code>true</code> ).
------	---

`constructNextBatch` performs the following steps:

1. Requesting the latest offsets from every streaming source (of the streaming query)
2. Updating `availableOffsets StreamProgress` with the latest available offsets
3. Updating batch metadata with the current event-time watermark and batch timestamp
4. Checking whether to construct the next micro-batch or not (skip it)

In the end, `constructNextBatch` returns whether the next streaming micro-batch was constructed or skipped.

Note	<code>constructNextBatch</code> is used exclusively when <code>MicroBatchExecution</code> is requested to run the activated streaming query.
------	--

## Requesting Latest Offsets from Streaming Sources (`getOffset`, `setOffsetRange` and `getEndOffset` Phases)

`constructNextBatch` firstly requests every `streaming source` for the latest offsets.

Note	<code>constructNextBatch</code> checks out the latest offset in every streaming data source sequentially, i.e. one data source at a time.
------	---

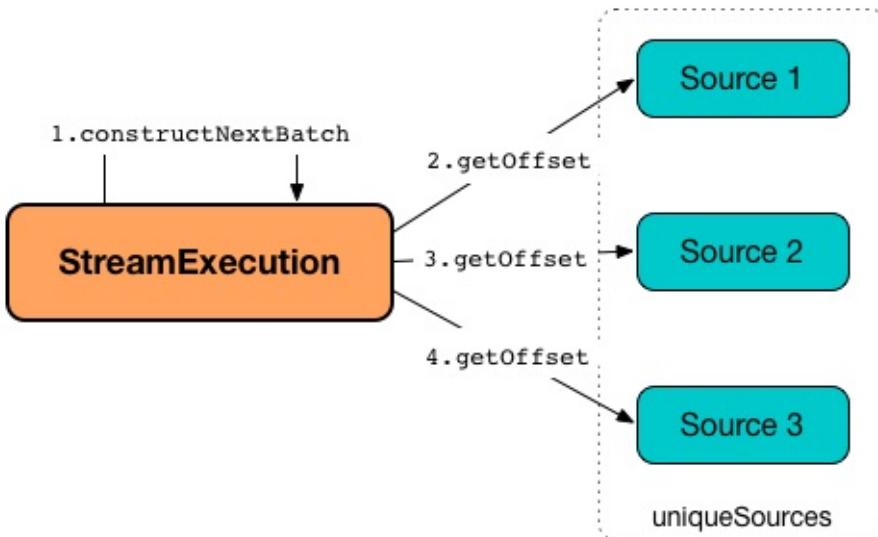


Figure 1. MicroBatchExecution's Getting Offsets From Streaming Sources

For every **streaming source** (Data Source API V1), `constructNextBatch` **updates the status message** to the following:

```
Getting offsets from [source]
```

In **getOffset time-tracking section**, `constructNextBatch` **requests the source for the latest offset**.

For every **MicroBatchReader** (Data Source API V2), `constructNextBatch` **updates the status message** to the following:

```
Getting offsets from [source]
```

In **setOffsetRange time-tracking section**, `constructNextBatch` **finds the available offsets of the source (in the available offset internal registry) and, if found, requests the MicroBatchReader to deserialize the offset (from JSON format)**. `constructNextBatch` **requests the MicroBatchReader to set the desired offset range**.

In **getEndOffset time-tracking section**, `constructNextBatch` **requests the MicroBatchReader for the end offset**.

## Updating availableOffsets StreamProgress with Latest Available Offsets

`constructNextBatch` **updates the availableOffsets StreamProgress with the latest reported offsets**.

## Updating Batch Metadata with Current Event-Time Watermark and Batch Timestamp

`constructNextBatch` updates the [batch metadata](#) with the current [event-time watermark](#) (from the [WatermarkTracker](#)) and the batch timestamp.

## Checking Whether to Construct Next Micro-Batch or Not (Skip It)

`constructNextBatch` checks whether or not the next streaming micro-batch should be constructed (`lastExecutionRequiresAnotherBatch`).

`constructNextBatch` uses the [last IncrementalExecution](#) if the [last execution requires another micro-batch](#) (using the [batch metadata](#)) and the given `noDataBatchesEnabled` flag is enabled (`true`).

`constructNextBatch` also [checks out whether new data is available](#) (based on [available](#) and [committed offsets](#)).

Note	<code>shouldConstructNextBatch</code> local flag is enabled ( <code>true</code> ) when <a href="#">there is new data available (based on offsets)</a> or the <a href="#">last execution requires another micro-batch</a> (and the given <code>noDataBatchesEnabled</code> flag is enabled).
------	---

`constructNextBatch` prints out the following TRACE message to the logs:

```
noDataBatchesEnabled = [noDataBatchesEnabled],  
lastExecutionRequiresAnotherBatch =  
[lastExecutionRequiresAnotherBatch], isNewDataAvailable =  
[isNewDataAvailable], shouldConstructNextBatch =  
[shouldConstructNextBatch]
```

`constructNextBatch` branches off per whether to [constructs](#) or [skip](#) the next batch (per `shouldConstructNextBatch` flag in the above TRACE message).

## Constructing Next Micro-Batch — `shouldConstructNextBatch` Flag Enabled

With the `shouldConstructNextBatch` flag enabled (`true`), `constructNextBatch` [updates the status message](#) to the following:

```
Writing offsets to log
```

In `walCommit` time-tracking section, `constructNextBatch` requests the `availableOffsets` `StreamProgress` to convert to `OffsetSeq` (with the `BaseStreamingSources` and the `current batch metadata (event-time watermark and timestamp)`) that is in turn added to the `write-ahead log` for the `current batch ID`.

`constructNextBatch` prints out the following INFO message to the logs:

```
Committed offsets for batch [currentBatchId]. Metadata [offsetSeqMetadata]
```

Note	<code>FIXME ( if (currentBatchId != 0) ... )</code>
------	---

Note	<code>FIXME ( if (minLogEntriesToMaintain &lt; currentBatchId) ... )</code>
------	---

`constructNextBatch` turns the `noNewData` internal flag off (`false`).

In case of a failure while adding the available offsets to the write-ahead log,

`constructNextBatch` throws an `AssertionError`:

```
Concurrent update to the log. Multiple streaming jobs detected for [currentBatchId]
```

## Skipping Next Micro-Batch — `shouldConstructNextBatch` Flag Disabled

With the `shouldConstructNextBatch` flag disabled (`false`), `constructNextBatch` turns the `noNewData` flag on (`true`) and wakes up (*notifies*) all threads waiting for the `awaitProgressLockCondition` lock.

## Running Single Streaming Micro-Batch — `runBatch` Internal Method

```
runBatch(  
    sparkSessionToRunBatch: SparkSession): Unit
```

`runBatch` prints out the following DEBUG message to the logs (with the `current batch ID`):

```
Running batch [currentBatchId]
```

`runBatch` then performs the following steps (aka *phases*):

1. `getBatch` Phase — Creating Logical Query Plans For Unprocessed Data From Sources and `MicroBatchReaders`

2. Transforming Logical Plan to Include Sources and MicroBatchReaders with New Data
3. Transforming CurrentTimestamp and CurrentDate Expressions (Per Batch Metadata)
4. Adapting Transformed Logical Plan to Sink with StreamWriteSupport
5. Setting Local Properties
6. queryPlanning Phase — Creating and Preparing IncrementalExecution for Execution
7. nextBatch Phase — Creating DataFrame (with IncrementalExecution for New Data)
8. addBatch Phase — Adding DataFrame With New Data to Sink
9. Updating Watermark and Committing Offsets to Offset Commit Log

In the end, `runBatch` prints out the following DEBUG message to the logs (with the `current batch ID`):

```
Completed batch [currentBatchId]
```

Note

`runBatch` is used exclusively when `MicroBatchExecution` is requested to run an activated streaming query (and there is new data to process).

## getBatch Phase — Creating Logical Query Plans For Unprocessed Data From Sources and MicroBatchReaders

In `getBatch` time-tracking section, `runBatch` goes over the `available offsets` and processes every `Source` and `MicroBatchReader` (associated with the available offsets) to create logical query plans (`newData`) for data processing (per offset ranges).

Note

`runBatch` requests sources and readers for data per offset range sequentially, one by one.

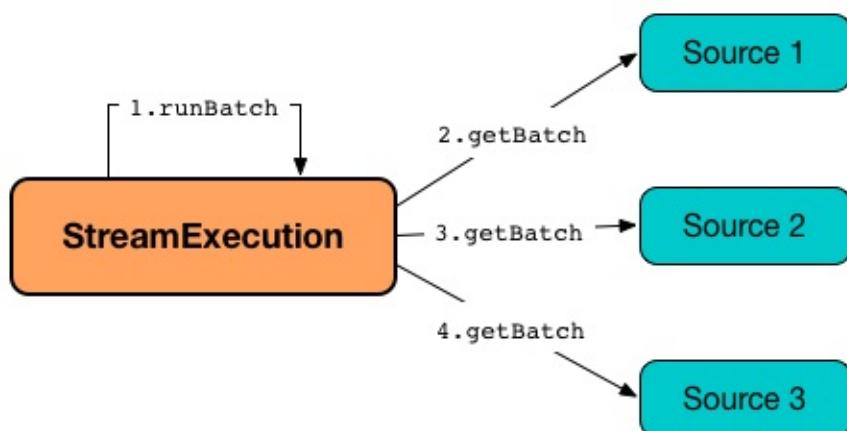


Figure 2. StreamExecution's Running Single Streaming Batch (getBatch Phase)

## getBatch Phase and Sources

For a `Source` (with the available `offsets` different from the `committedOffsets` registry), `runBatch` does the following:

- Requests the `committedOffsets` for the committed offsets for the `Source` (if available)
- Requests the `Source` for a `dataframe for the offset range` (the current and available offsets)

`runBatch` prints out the following DEBUG message to the logs.

```
Retrieving data from [source]: [current] -> [available]
```

In the end, `runBatch` returns the `Source` and the logical plan of the streaming dataset (for the offset range).

In case the `Source` returns a dataframe that is not streaming, `runBatch` throws an `AssertionError`:

```
DataFrame returned by getBatch from [source] did not have isStreaming=true\n[logicalQueryPlan]
```

## getBatch Phase and MicroBatchReaders

For a `MicroBatchReader` (with the available `offsets` different from the `committedOffsets` registry), `runBatch` does the following:

- Requests the `committedOffsets` for the committed offsets for the `MicroBatchReader` (if available)
- Requests the `MicroBatchReader` to `deserialize the committed offsets` (if available)
- Requests the `MicroBatchReader` to `deserialize the available offsets` (only for `SerializedOffsets`)
- Requests the `MicroBatchReader` to `set the offset range` (the current and available offsets)

`runBatch` prints out the following DEBUG message to the logs.

```
Retrieving data from [reader]: [current] -> [availableV2]
```

`runBatch` looks up the `DataSourceV2` and the options for the `MicroBatchReader` (in the `readerToDataSourceMap` internal registry).

In the end, `runBatch` requests the `MicroBatchReader` for the `read schema` and creates a `StreamingDataSourceV2Relation` logical operator (with the read schema, the `DataSourceV2`, options, and the `MicroBatchReader` ).

## Transforming Logical Plan to Include Sources and MicroBatchReaders with New Data

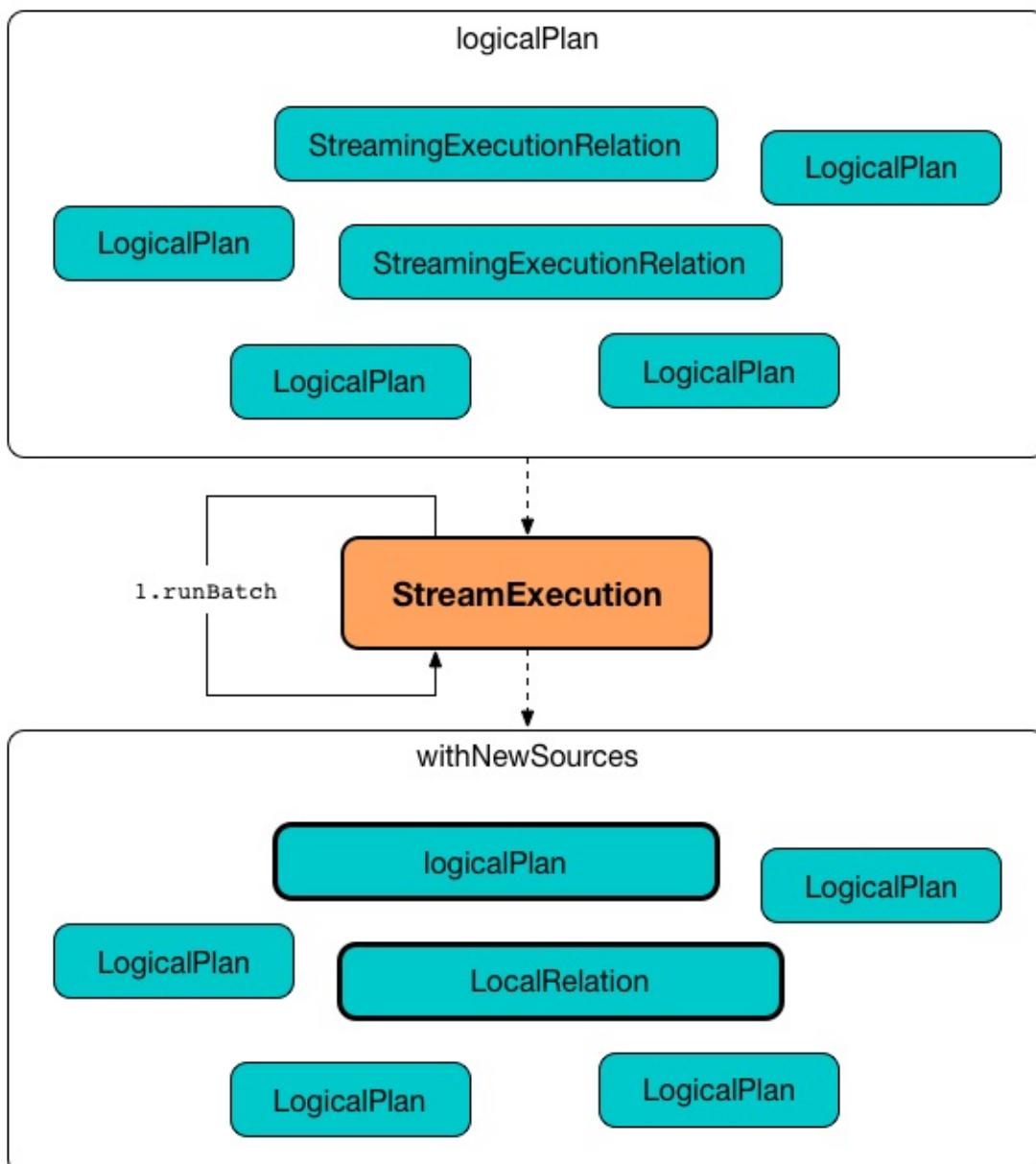


Figure 3. StreamExecution's Running Single Streaming Batch (and Transforming Logical Plan for New Data)

`runBatch` transforms the `analyzed logical plan` to include `Sources and MicroBatchReaders with new data` (`newBatchesPlan` with logical plans to process data that has arrived since the last batch).

For every `StreamingExecutionRelation` (with a `Source` or `MicroBatchReader`), `runBatch` tries to find the corresponding logical plan for processing new data.

**Note**

`StreamingExecutionRelation` logical operator is used to represent a streaming source or reader in the `logical query plan` (of a streaming query).

If the logical plan is found, `runBatch` makes the plan a child operator of `Project` (with `Aliases`) logical operator and replaces the `StreamingExecutionRelation`.

Otherwise, if not found, `runBatch` simply creates an empty streaming `LocalRelation` (for scanning data from an empty local collection).

In case the number of columns in dataframes with new data and `StreamingExecutionRelation`'s do not match, `runBatch` throws an `AssertionError`:

```
Invalid batch: [output] != [dataPlan.output]
```

## Transforming CurrentTimestamp and CurrentDate Expressions (Per Batch Metadata)

`runBatch` replaces all `CurrentTimestamp` and `CurrentDate` expressions in the transformed logical plan (with new data) with the `current batch timestamp` (based on the `batch metadata`).

**Note**

`currentTimestamp` and `currentDate` expressions correspond to `current_timestamp` and `current_date` standard function, respectively.

Read up [The Internals of Spark SQL](#) to learn more about the standard functions.

## Adapting Transformed Logical Plan to Sink with StreamWriteSupport

`runBatch` adapts the transformed logical plan (with new data and current batch timestamp) for the new `StreamWriteSupport` sinks (per the type of the `BaseStreamingSink`).

For a `StreamWriteSupport` (Data Source API V2), `runBatch` requests the `StreamWriteSupport` for a `StreamWriter` (for the `runId`, the output schema, the `OutputMode`, and the `extra options`). `runBatch` then creates a `WriteToDataSourceV2` logical operator with a new `MicroBatchWriter` as a child operator (for the `current batch ID` and the `StreamWriter`).

For a `Sink` (Data Source API V1), `runBatch` changes nothing.

For any other `BaseStreamingSink` type, `runBatch` simply throws an `IllegalArgumentException`:

```
unknown sink type for [sink]
```

## Setting Local Properties

`runBatch` sets the [local properties](#).

Table 1. `runBatch`'s Local Properties

Local Property	Value
<code>streaming.sql.batchId</code>	<code>currentBatchId</code>
<code>__is_continuous_processing</code>	<code>false</code>

## queryPlanning Phase — Creating and Preparing IncrementalExecution for Execution

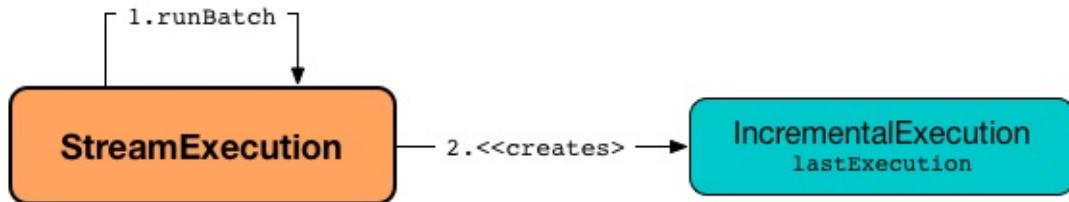


Figure 4. StreamExecution's Query Planning (queryPlanning Phase)

In [queryPlanning time-tracking section](#), `runBatch` creates a new `IncrementalExecution` with the following:

- Transformed logical plan
- Output mode
- `state checkpoint directory`
- Run id
- Batch id
- Batch Metadata (Event-Time Watermark and Timestamp)

In the end (of the `queryPlanning phase`), `runBatch` requests the `IncrementalExecution` to prepare the transformed logical plan for execution (i.e. execute the `executedPlan` query execution phase).

Tip

Read up on the `executedPlan` query execution phase in [The Internals of Spark SQL](#).

## nextBatch Phase — Creating DataFrame (with IncrementalExecution for New Data)

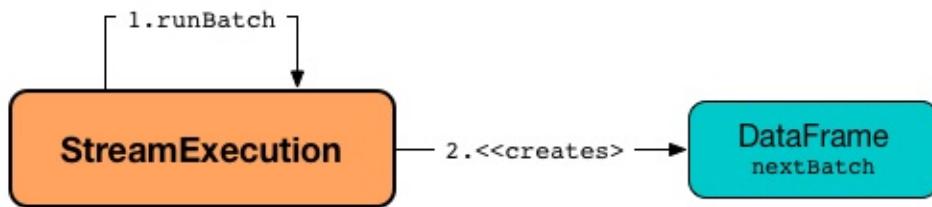


Figure 5. StreamExecution Creates DataFrame with New Data

`runBatch` creates a new `DataFrame` with the new [IncrementalExecution](#).

The `DataFrame` represents the result of executing the current micro-batch of the streaming query.

## addBatch Phase — Adding DataFrame With New Data to Sink

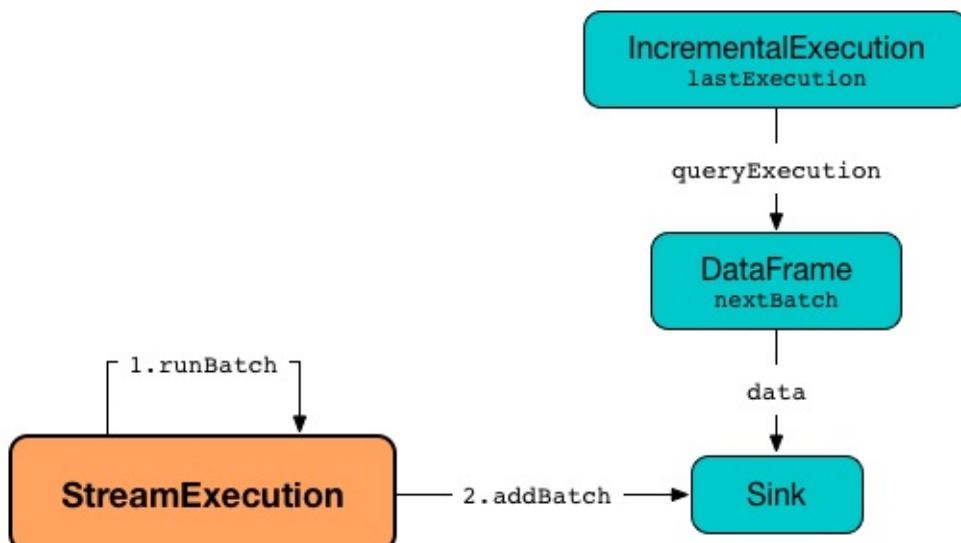


Figure 6. StreamExecution Adds DataFrame With New Data to Sink

In [addBatch time-tracking section](#), `runBatch` adds the `DataFrame` with new data to the [BaseStreamingSink](#).

For a [Sink](#) (Data Source API V1), `runBatch` simply requests the `Sink` to [add the DataFrame](#) (with the batch ID).

For a [StreamWriterSupport](#) (Data Source API V2), `runBatch` simply requests the `DataFrame` with new data to collect (which simply forces execution of the [MicroBatchWriter](#)).

Note	<code>runBatch</code> uses <code>SQLExecution.withNewExecutionId</code> to execute and track all the Spark jobs under one execution id (so it is reported as one single multi-job execution, e.g. in web UI).
------	---

**Note**

`SQLExecution.withNewExecutionId` posts a `SparkListenerSQLExecutionStart` event before execution and a `SparkListenerSQLExecutionEnd` event right afterwards.

**Tip**

Register `SparkListener` to get notified about the SQL execution events (`SparkListenerSQLExecutionStart` and `SparkListenerSQLExecutionEnd`).

Read up on `SparkListener` in [The Internals of Apache Spark](#).

## Updating Watermark and Committing Offsets to Offset Commit Log

`runBatch` requests the `WatermarkTracker` to update event-time watermark (with the `executedPlan` of the `IncrementalExecution`).

`runBatch` requests the `Offset Commit Log` to persisting metadata of the streaming micro-batch (with the current `batch ID` and event-time watermark of the `WatermarkTracker`).

In the end, `runBatch` adds the `available offsets` to the `committed offsets` (and updates the `offsets` of every `BaseStreamingSource` with new data in the current micro-batch).

## Stopping Stream Processing (Execution of Streaming Query) — `stop` Method

`stop(): Unit`

**Note**

`stop` is part of the `StreamingQuery Contract` to stop a streaming query.

`stop` sets the `state` to `TERMINATED`.

When the `stream execution thread` is alive, `stop` requests the current `sparkContext` to `cancelJobGroup` identified by the `runId` and waits for this thread to die. Just to make sure that there are no more streaming jobs, `stop` requests the current `sparkContext` to `cancelJobGroup` identified by the `runId` again.

In the end, `stop` prints out the following INFO message to the logs:

Query [prettyIdString] was stopped

## Checking Whether New Data Is Available (Based on Available and Committed Offsets)

### — `isNewDataAvailable` Internal Method

```
isNewDataAvailable: Boolean
```

`isNewDataAvailable` checks whether there is a streaming source (in the [available offsets](#)) for which [committed offsets](#) are different from the available offsets or not available (committed) at all.

`isNewDataAvailable` is positive (`true`) when there is at least one such streaming source.

Note

`isNewDataAvailable` is used when `MicroBatchExecution` is requested to [run an activated streaming query](#) and [construct the next streaming micro-batch](#).

## Analyzed Logical Plan With Unique StreamingExecutionRelation Operators — `logicalPlan` Lazy Property

```
logicalPlan: LogicalPlan
```

Note

`logicalPlan` is part of [StreamExecution Contract](#) to be the analyzed logical plan of the streaming query.

`logicalPlan` resolves ([replaces](#)) [StreamingRelation](#), [StreamingRelationV2](#) logical operators to [StreamingExecutionRelation](#) logical operators. `logicalPlan` uses the transformed logical plan to set the [uniqueSources](#) and [sources](#) internal registries to be the [BaseStreamingSources](#) of all the [StreamingExecutionRelations](#) unique and not, respectively.

Note

`logicalPlan` is a Scala lazy value and so the initialization is guaranteed to happen only once at the first access (and is cached for later use afterwards).

Internally, `logicalPlan` transforms the [analyzed logical plan](#).

For every [StreamingRelation](#) logical operator, `logicalPlan` tries to replace it with the [StreamingExecutionRelation](#) that was used earlier for the same [streamingRelation](#) (if used multiple times in the plan) or creates a new one. While creating a new [StreamingExecutionRelation](#), `logicalPlan` requests the [DataSource](#) to [create a streaming Source](#) with the metadata path as `sources/uniqueID` directory in the [checkpoint root directory](#). `logicalPlan` prints out the following INFO message to the logs:

```
Using Source [source] from DataSourceV1 named '[sourceName]' [dataSourceV1]
```

For every `StreamingRelationV2` logical operator with a `MicroBatchReadSupport` data source (which is not on the list of `spark.sql.streaming.disabledV2MicroBatchReaders`), `logicalPlan` tries to replace it with the `StreamingExecutionRelation` that was used earlier for the same `StreamingRelationV2` (if used multiple times in the plan) or creates a new one. While creating a new `StreamingExecutionRelation`, `logicalPlan` requests the `MicroBatchReadSupport` to create a `MicroBatchReader` with the metadata path as `sources/uniqueID` directory in the `checkpoint root directory`. `logicalPlan` prints out the following INFO message to the logs:

```
Using MicroBatchReader [reader] from DataSourceV2 named '[sourceName]' [dataSourceV2]
```

For every other `StreamingRelationV2` logical operator, `logicalPlan` tries to replace it with the `StreamingExecutionRelation` that was used earlier for the same `StreamingRelationV2` (if used multiple times in the plan) or creates a new one. While creating a new `StreamingExecutionRelation`, `logicalPlan` requests the `StreamingRelation` for the underlying `DataSource` that is in turn requested to create a streaming Source with the metadata path as `sources/uniqueID` directory in the `checkpoint root directory`. `logicalPlan` prints out the following INFO message to the logs:

```
Using Source [source] from DataSourceV2 named '[sourceName]' [dataSourceV2]
```

`logicalPlan` requests the transformed analyzed logical plan for all `StreamingExecutionRelations` that are then requested for `BaseStreamingSources`, and saves them as the `sources` internal registry.

In the end, `logicalPlan` sets the `uniqueSources` internal registry to be the unique `BaseStreamingSources` above.

`logicalPlan` throws an `AssertionError` when not executed on the `stream execution thread`.

```
logicalPlan must be initialized in QueryExecutionThread but the current thread was [currentThread]
```

## streaming.sql.batchId Local Property

`MicroBatchExecution` defines `streaming.sql.batchId` as the name of the local property to be the current **batch** or **epoch IDs** (that Spark tasks can use)

`streaming.sql.batchId` is used when:

- `MicroBatchExecution` is requested to run a single streaming micro-batch (and sets the property to be the current batch ID)
- `DataWritingSparkTask` is requested to run (and needs an epoch ID)

## Internal Properties

Name	Description		
<code>isCurrentBatchConstructed</code>	<p>Flag to control whether to run a streaming micro-batch (<code>true</code>) or not (<code>false</code>)</p> <p>Default: <code>false</code></p> <ul style="list-style-type: none"> <li>• When disabled (<code>false</code>), changed to whatever constructing the next streaming micro-batch gives back when running activated streaming query</li> <li>• Disabled (<code>false</code>) after running a streaming micro-batch (when enabled after constructing the next streaming micro-batch)</li> <li>• Enabled (<code>true</code>) when populating start offsets (when running an activated streaming query) and restarting a streaming query from a checkpoint (using the Offset Write-Ahead Log)</li> <li>• Disabled (<code>false</code>) when populating start offsets (when running an activated streaming query) and restarting a streaming query from a checkpoint when the latest offset checkpointed (written) to the offset write-ahead log has been successfully processed and committed to the Offset Commit Log</li> </ul>		
<code>readerToDataSourceMap</code>	<p>( <code>Map[MicroBatchReader, (DataSourceV2, Map[String, String])]</code> )</p>		
<code>sources</code>	<p>Streaming sources and readers (of the <code>StreamingExecutionRelations</code> of the analyzed logical query plan of the streaming query)</p> <p>Default: (empty)</p> <table border="1"> <tr> <td>Note</td><td> <p><code>sources</code> is part of the ProgressReporter Contract for the streaming sources of the streaming query.</p> </td></tr> </table> <ul style="list-style-type: none"> <li>• Initialized when <code>MicroBatchExecution</code> is requested for the transformed logical query plan</li> </ul> <p>Used when:</p>	Note	<p><code>sources</code> is part of the ProgressReporter Contract for the streaming sources of the streaming query.</p>
Note	<p><code>sources</code> is part of the ProgressReporter Contract for the streaming sources of the streaming query.</p>		

	<ul style="list-style-type: none"><li>• Populating start offsets (for the available and committed offsets)</li><li>• Constructing or skipping next streaming micro-batch (and persisting offsets to write-ahead log)</li></ul>
watermarkTracker	<p>WatermarkTracker that is created when MicroBatchExecution is requested to populate start offsets (when requested to run an activated streaming query)</p>

# MicroBatchWriter — Data Source Writer in Micro-Batch Stream Processing (Data Source API V2)

`MicroBatchWriter` is a `DataSourceWriter` (Spark SQL) that uses the given batch ID as the epoch when requested to commit, abort and create a `WriterFactory` for a given [StreamWriter](#) in [Micro-Batch Stream Processing](#).

Tip

Read up on [DataSourceWriter](#) in [The Internals of Spark SQL](#) book.

`MicroBatchWriter` is part of the novel Data Source API V2 in Spark SQL.

`MicroBatchWriter` is created exclusively when `MicroBatchExecution` is requested to run a [streaming batch](#) (with a [StreamWriterSupport](#) streaming sink).

# MicroBatchReadSupport Contract — Data Sources with MicroBatchReaders

`MicroBatchReadSupport` is the extension of the `DataSourceV2` for data sources with a `MicroBatchReader` for Micro-Batch Stream Processing.

`MicroBatchReadSupport` defines a single `createMicroBatchReader` method to create a `MicroBatchReader`.

```
MicroBatchReader createMicroBatchReader(
    Optional<StructType> schema,
    String checkpointLocation,
    DataSourceOptions options)
```

`createMicroBatchReader` is used when:

- `MicroBatchExecution` is requested for the analyzed logical plan (and creates a `StreamingExecutionRelation` for a `StreamingRelationV2` with a `MicroBatchReadSupport` data source)
- `DataStreamReader` is requested to create a streaming query for a `MicroBatchReadSupport` data source

Table 1. MicroBatchReadSupports

MicroBatchReadSupport	Description
KafkaSourceProvider	Data source provider for <code>kafka</code> format
RateStreamProvider	Data source provider for <code>rate</code> format
TextSocketSourceProvider	Data source provider for <code>socket</code> format

# MicroBatchReader Contract — Data Source Readers in Micro-Batch Stream Processing (Data Source API V2)

`MicroBatchReader` is the [extension](#) of Spark SQL's `DataSourceReader` (and `BaseStreamingSource`) contracts for [data source readers](#) in [Micro-Batch Stream Processing](#).

`MicroBatchReader` is part of the novel Data Source API V2 in Spark SQL.

Tip

Read up on [Data Source API V2](#) in [The Internals of Spark SQL](#) book.

Table 1. MicroBatchReader Contract

Method	Description
commit	<pre>void commit(<code>offset</code> end)</pre> <p>Used when...FIXME</p>
deserializeOffset	<pre>offset deserializeOffset(<code>String</code> json)</pre> <p>Deserializes <code>offset</code> (from JSON format)</p> <p>Used when...FIXME</p>
getEndOffset	<pre>offset getEndOffset()</pre> <p>End <code>offset</code> of this reader</p> <p>Used when...FIXME</p>
getStartOffset	<pre>offset getStartOffset()</pre> <p>Start (beginning) <code>offsets</code> of this reader</p> <p>Used when...FIXME</p>
setOffsetRange	<pre>void setOffsetRange(     <code>Optional&lt;Offset&gt;</code> start,     <code>Optional&lt;Offset&gt;</code> end)</pre> <p>Sets the desired offset range for input partitions created from this reader (for data scan)</p> <p>Used when...FIXME</p>

Table 2. MicroBatchReaders

MicroBatchReader	Description
KafkaMicroBatchReader	
MemoryStream	
RateStreamMicroBatchReader	
TextSocketMicroBatchReader	



# WatermarkTracker

`WatermarkTracker` tracks the event-time watermark of a streaming query (across `EventTimeWatermarkExec` operators in a physical query plan) based on a given `MultipleWatermarkPolicy`.

`WatermarkTracker` is used exclusively in `MicroBatchExecution`.

`WatermarkTracker` is created (using the `factory method`) when `MicroBatchExecution` is requested to `populate start offsets` (when requested to `run an activated streaming query`).

`WatermarkTracker` takes a single `MultipleWatermarkPolicy` to be created.

`MultipleWatermarkPolicy` can be one of the following:

- `MaxWatermark` (alias: `min`)
- `MinWatermark` (alias: `max`)

Enable `ALL` logging level for `org.apache.spark.sql.execution.streaming.WatermarkTracker` to see what happens inside.

Tip

Add the following line to `conf/log4j.properties`:

```
log4j.logger.org.apache.spark.sql.execution.streaming.WatermarkTracker=ALL
```

Refer to [Logging](#).

## Creating WatermarkTracker — apply Factory Method

```
apply(conf: RuntimeConfig): WatermarkTracker
```

`apply` uses the `spark.sql.streaming.multipleWatermarkPolicy` configuration property for the global watermark policy (default: `min`) and creates a `WatermarkTracker`.

Note

`apply` is used exclusively when `MicroBatchExecution` is requested to `populate start offsets` (when requested to `run an activated streaming query`).

## setWatermark Method

```
setWatermark(newWatermarkMs: Long): Unit
```

`setWatermark` simply updates the [global event-time watermark](#) to the given `newWatermarkMs`.

**Note**

`setWatermark` is used exclusively when `MicroBatchExecution` is requested to [populate start offsets](#) (when requested to [run an activated streaming query](#)).

## Updating Event-Time Watermark — `updateWatermark` Method

`updateWatermark(executedPlan: SparkPlan): Unit`

`updateWatermark` requests the given physical operator (`SparkPlan`) to collect all [EventTimeWatermarkExec](#) unary physical operators.

`updateWatermark` simply exits when no `EventTimeWatermarkExec` was found.

`updateWatermark` ...FIXME

**Note**

`updateWatermark` is used exclusively when `MicroBatchExecution` is requested to [run a single streaming batch](#) (when requested to [run an activated streaming query](#)).

## Internal Properties

Name	Description
<code>globalWatermarkMs</code>	<p>Current <b>global event-time watermark</b> per <a href="#">MultipleWatermarkPolicy</a> (across <a href="#">all EventTimeWatermarkExec operators</a> in a physical query plan)</p> <p>Default: <code>0</code></p> <p>Used when...FIXME</p>
<code>operatorToWatermarkMap</code>	<p>Event-time watermark per <a href="#">EventTimeWatermarkExec</a> physical operator (<code>mutable.HashMap[Int, Long]</code>)</p> <p>Used when...FIXME</p>

# Source Contract — Streaming Sources for Micro-Batch Stream Processing (Data Source API V1)

`Source` is the [extension](#) of the `BaseStreamingSource` contract for [streaming sources](#) that work with "continuous" stream of data identified by [offsets](#).

`Source` is part of Data Source API V1 and used in [Micro-Batch Stream Processing](#) only.

For fault tolerance, `Source` must be able to replay an arbitrary sequence of past data in a stream using a range of offsets. This is the assumption so Structured Streaming can achieve end-to-end exactly-once guarantees.

Table 1. Source Contract

Method	Description
<code>commit</code>	<pre>commit(end: Offset): Unit</pre> <p>Commits data up to the end <a href="#">offset</a>, i.e. informs the source that Spark has completed processing all data for offsets less than or equal to the end offset and will only request offsets greater than the end offset in the future.</p> <p>Used exclusively when <a href="#">MicroBatchExecution</a> stream execution engine (<a href="#">Micro-Batch Stream Processing</a>) is requested to <a href="#">write offsets to a commit log (walCommit phase)</a> while <a href="#">running an activated streaming query</a>.</p>
<code>getBatch</code>	<pre>getBatch(   start: Option[Offset],   end: Offset): DataFrame</pre> <p>Generating a streaming <code>DataFrame</code> with data between the start and end <a href="#">offsets</a></p> <p>Start offset can be undefined (<code>None</code>) to indicate that the batch should begin with the first record</p> <p>Used when <a href="#">MicroBatchExecution</a> stream execution engine (<a href="#">Micro-Batch Stream Processing</a>) is requested to <a href="#">run an activated streaming query</a>, namely:</p> <ul style="list-style-type: none"> <li>• Populate start offsets from checkpoint (resuming from checkpoint)</li> <li>• Request unprocessed data from all sources (getBatch phase)</li> </ul>

	<p><code>getOffset: Option[Offset]</code></p> <p><code>getOffset</code></p> <p>Latest (maximum) <a href="#">offset</a> of the source (or <code>None</code> to denote no data)</p> <p>Used exclusively when <a href="#">MicroBatchExecution</a> stream execution engine (<a href="#">Micro-Batch Stream Processing</a>) is requested for <a href="#">latest offsets of all sources (getOffset phase)</a> while <a href="#">running activated streaming query</a>.</p>		
<code>schema</code>	<p><code>schema: StructType</code></p> <p>Schema of the source</p> <table border="1"> <tr> <td>Note</td><td><code>schema</code> <i>seems</i> to be used for tests only and a duplication of <a href="#">StreamSourceProvider.sourceSchema</a>.</td></tr> </table>	Note	<code>schema</code> <i>seems</i> to be used for tests only and a duplication of <a href="#">StreamSourceProvider.sourceSchema</a> .
Note	<code>schema</code> <i>seems</i> to be used for tests only and a duplication of <a href="#">StreamSourceProvider.sourceSchema</a> .		

Table 2. Sources

Source	Description
<a href="#">FileStreamSource</a>	Part of file-based data sources ( <code>FileFormat</code> )
<a href="#">KafkaSource</a>	Part of <a href="#">kafka</a> data source

# StreamSourceProvider Contract — Streaming Source Providers for Micro-Batch Stream Processing (Data Source API V1)

`StreamSourceProvider` is the contract of data source providers that can create a streaming source for a format (e.g. text file) or system (e.g. Apache Kafka).

`StreamSourceProvider` is part of Data Source API V1 and used in Micro-Batch Stream Processing only.

Table 1. StreamSourceProvider Contract

Method	Description
<code>createSource</code>	<pre>createSource(     sqlContext: SQLContext,     metadataPath: String,     schema: Option[StructType],     providerName: String,     parameters: Map[String, String]): Source</pre> <p>Creates a <a href="#">streaming source</a></p> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 10px;"> <p><b>Note</b> <code>metadataPath</code> is the value of the optional user-specified <code>checkpointLocation</code> option or resolved by <a href="#">StreamingQueryManager</a>.</p> </div> <p>Used exclusively when <code>DataSource</code> is requested to <a href="#">create a streaming source</a> (when <code>MicroBatchExecution</code> is requested to <a href="#">initialize the analyzed logical plan</a>)</p>
<code>sourceSchema</code>	<pre>sourceSchema(     sqlContext: SQLContext,     schema: Option[StructType],     providerName: String,     parameters: Map[String, String]): (String, StructType)</pre> <p>The name and schema of the <a href="#">streaming source</a></p> <p>Used exclusively when <code>DataSource</code> is requested for <a href="#">metadata of a streaming source</a> (when <code>MicroBatchExecution</code> is requested to <a href="#">initialize the analyzed logical plan</a>)</p>
<b>Note</b>	<p><a href="#">KafkaSourceProvider</a> is the only known <code>StreamSourceProvider</code> in Spark Structured Streaming.</p>



# Sink Contract — Streaming Sinks for Micro-Batch Stream Processing

`Sink` is the extension of the [BaseStreamingSink contract](#) for streaming sinks that can add batches to an output.

`Sink` is part of Data Source API V1 and used in [Micro-Batch Stream Processing](#) only.

Table 1. Sink Contract

Method	Description
<code>addBatch</code>	<pre>addBatch(     batchId: Long,     data: DataFrame): Unit</pre> <p>Adds a batch of data to the sink</p> <p>Used exclusively when <a href="#">MicroBatchExecution</a> stream execution engine (<a href="#">Micro-Batch Stream Processing</a>) is requested to <a href="#">add a streaming batch to a sink (addBatch phase)</a> while <a href="#">running an activated streaming query</a>.</p>

Table 2. Sinks

Sink	Description
<a href="#">FileStreamSink</a>	Used in file-based data sources ( <a href="#">FileFormat</a> )
<a href="#">ForeachBatchSink</a>	Used for <a href="#">DataStreamWriter.foreachBatch</a> streaming operator
<a href="#">KafkaSink</a>	Used for <a href="#">kafka</a> output format
<a href="#">MemorySink</a>	Used for <code>memory</code> output format

# StreamSinkProvider Contract

`StreamSinkProvider` is the abstraction of providers that can create a streaming sink for a file format (e.g. `parquet`) or system (e.g. `kafka`).

**Important**

`StreamWriterSupport` is a newer version of `StreamSinkProvider` (aka `DataSource API V2`) and new data sources should use the contract instead.

Table 1. StreamSinkProvider Contract

Method	Description
<code>createSink</code>	<pre>createSink(     sqlContext: SQLContext,     parameters: Map[String, String],     partitionColumns: Seq[String],     outputMode: OutputMode): Sink</pre> <p>Creates a <a href="#">streaming sink</a></p> <p>Used exclusively when <code>DataSource</code> is requested for a <a href="#">streaming sink</a> (when <code>DataStreamWriter</code> is requested to <a href="#">start a streaming query</a>)</p>

**Note**

`KafkaSourceProvider` is the only known `StreamSinkProvider` in Spark Structured Streaming.

# Continuous Stream Processing (Structured Streaming V2)

**Continuous Stream Processing** is one of the two stream processing engines in [Spark Structured Streaming](#) that is used for execution of structured streaming queries with [Trigger.Continuous](#) trigger.

Note

The other feature-richer stream processing engine is [Micro-Batch Stream Processing](#).

Continuous Stream Processing execution engine uses the novel **Data Source API V2** (Spark SQL) and for the very first time makes stream processing truly **continuous**.

Tip

Read up on [Data Source API V2](#) in [The Internals of Spark SQL](#) book.

Because of the two innovative changes Continuous Stream Processing is often referred to as **Structured Streaming V2**.

```
import org.apache.spark.sql.streaming.Trigger
import scala.concurrent.duration._

val sq = spark
  .readStream
  .format("rate")
  .load
  .writeStream
  .format("console")
  .option("truncate", false)
  .trigger(Trigger.Continuous(15.seconds)) // <-- Uses ContinuousExecution for execution
  .queryName("rate2console")
  .start

scala> :type sq
org.apache.spark.sql.streaming.StreamingQuery

assert(sq.isActive)

// sq.stop
```

Under the covers, Continuous Stream Processing uses [ContinuousExecution](#) stream execution engine. When requested to [run an activated streaming query](#), [ContinuousExecution](#) adds [WriteToContinuousDataSourceExec](#) physical operator as the top-level operator in the physical query plan of the streaming query.

```

scala> :type sq
org.apache.spark.sql.streaming.StreamingQuery

scala> sq.explain
== Physical Plan ==
WriteToContinuousDataSource ConsoleWriter[numRows=20, truncate=false]
+- *(1) Project [timestamp#758, value#759L]
  +- *(1) ScanV2 rate[timestamp#758, value#759L]

```

From now on, you may think of a streaming query as a soon-to-be-generated [ContinuousWriteRDD](#) - an RDD data structure that Spark developers use to describe a distributed computation.

When the streaming query is started (and the top-level `writeToContinuousDataSourceExec` physical operator is requested to [execute and generate a recipe for a distributed computation \(as an `RDD\[InternalRow\]`\)](#)), it simply requests the underlying `ContinuousWriteRDD` to collect.

That collect operator is how a Spark job is run (as tasks over all partitions of the RDD) as described by the [ContinuousWriteRDD.compute](#) "protocol" (a recipe for the tasks to be scheduled to run on Spark executors).

Job Id (Job Group)	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0 (5b9edde1-e8cb-4ed5-a487-e841af09eba6)	rate2console id = b3074d1c-adbd-456f-a730-8aa0fc5bfc02 runId = 5b9edde1-e8cb-4ed5-a487-e841af09eba6 batch = init start at <>console>:33	2019/06/03 11:13:27	2.3 min	0/1	0/5 (5 running)

Figure 1. Creating Instance of StreamExecution

While the [tasks are computing partitions](#) (of the `ContinuousWriteRDD`), they keep running [until killed or completed](#). And that's the *ingenious design trick* of how the streaming query (as a Spark job with the distributed tasks running on executors) runs continuously and indefinitely.

When `DataStreamReader` is requested to [create a streaming query for a `ContinuousReadSupport` data source](#), it creates...FIXME

# ContinuousExecution — Stream Execution Engine of Continuous Stream Processing

`ContinuousExecution` is the [stream execution engine](#) of [Continuous Stream Processing](#).

`ContinuousExecution` is [created](#) when `StreamingQueryManager` is requested to [create a streaming query](#) with a `StreamWriterSupport` sink and a `ContinuousTrigger` (when `DataStreamWriter` is requested to [start an execution of the streaming query](#)).

`ContinuousExecution` can only run streaming queries with `StreamingRelationV2` with `ContinuousReadSupport` data source.

`ContinuousExecution` supports one `ContinuousReader` only in a [streaming query](#) (and asserts it when [addOffset](#) and [committing an epoch](#)). When requested for available [streaming sources](#), `ContinuousExecution` simply gives the [single ContinuousReader](#).

```
import org.apache.spark.sql.streaming.Trigger
import scala.concurrent.duration._

val sq = spark
  .readStream
  .format("rate")
  .load
  .writeStream
  .format("console")
  .option("truncate", false)
  .trigger(Trigger.Continuous(1.minute)) // <-- Gives ContinuousExecution
  .queryName("rate2console")
  .start

import org.apache.spark.sql.streaming.StreamingQuery
assert(sq.isInstanceOf[StreamingQuery])

// The following gives access to the internals
// And to ContinuousExecution
import org.apache.spark.sql.execution.streaming.StreamingQueryWrapper
val engine = sq.asInstanceOf[StreamingQueryWrapper].streamingQuery
import org.apache.spark.sql.execution.streaming.StreamExecution
assert(engine.isInstanceOf[StreamExecution])

import org.apache.spark.sql.execution.streaming.continuous.ContinuousExecution
val continuousEngine = engine.asInstanceOf[ContinuousExecution]
assert(continuousEngine.trigger == Trigger.Continuous(1.minute))
```

When [created](#) (for a streaming query), `ContinuousExecution` is given the [analyzed logical plan](#). The analyzed logical plan is immediately transformed to include a `ContinuousExecutionRelation` for every `StreamingRelationV2` with `ContinuousReadSupport` data source (and is the [logical plan](#) internally).

**Note**

`ContinuousExecution` uses the same instance of `continuousExecutionRelation` for the same instances of `StreamingRelationV2` with `ContinuousReadSupport` data source.

When requested to [run the streaming query](#), `ContinuousExecution` collects `ContinuousReadSupport` data sources (inside `ContinuousExecutionRelation`) from the analyzed logical plan and requests each and every `continuousReadSupport` to [create a `ContinuousReader`](#) (that are stored in `continuousSources` internal registry).

`ContinuousExecution` uses `__epoch_coordinator_id` local property for...FIXME

`ContinuousExecution` uses `__continuous_start_epoch` local property for...FIXME

`ContinuousExecution` uses `__continuous_epoch_interval` local property for...FIXME

**Tip**

Enable `ALL` logging level for `org.apache.spark.sql.execution.streaming.continuous.ContinuousExecution` to see what happens inside.

Add the following line to `conf/log4j.properties` :

```
log4j.logger.org.apache.spark.sql.execution.streaming.continuous.ContinuousExecution=INFO
```

Refer to [Logging](#).

## Running Activated Streaming Query

### — `runActivatedStream` Method

```
runActivatedStream(sparkSessionForStream: SparkSession): Unit
```

**Note**

`runActivatedStream` is part of [StreamExecution Contract](#) to run a streaming query.

`runActivatedStream` simply [runs the streaming query in continuous mode](#) as long as the state is [ACTIVE](#).

## Running Streaming Query in Continuous Mode

### — `runContinuous` Internal Method

```
runContinuous(sparkSessionForQuery: SparkSession): Unit
```

`runContinuous` initializes the `continuousSources` internal registry by traversing the `analyzed logical plan` to find `ContinuousExecutionRelation` leaf logical operators and requests their `ContinuousReadSupport` data sources to create a `ContinuousReader` (with the `sources` metadata directory under the `checkpoint` directory).

`runContinuous` initializes the `uniqueSources` internal registry to be the `continuousSources` distinct.

`runContinuous` gets the start offsets (they may or may not be available).

`runContinuous` transforms the `analyzed logical plan`. For every `ContinuousExecutionRelation` `runContinuous` finds the corresponding `ContinuousReader` (in the `continuousSources`), requests it to deserialize the start offsets (from their JSON representation), and then `setStartOffset`. In the end, `runContinuous` creates a `StreamingDataSourceV2Relation` (with the read schema of the `ContinuousReader` and the `ContinuousReader` itself).

`runContinuous` rewrites the transformed plan (with the `StreamingDataSourceV2Relation`) to use the new attributes from the source (the reader).

#### Note

`CurrentTimestamp` and `CurrentDate` expressions are not supported for continuous processing.

`runContinuous` requests the `StreamWriterSupport` to create a `StreamWriter` (with the run ID of the streaming query).

`runContinuous` creates a `WriteToContinuousDataSource` (with the `StreamWriter` and the transformed logical query plan).

`runContinuous` finds the only `ContinuousReader` (of the only `StreamingDataSourceV2Relation`) in the query plan with the `writeToContinuousDataSource`.

In `queryPlanning` time-tracking section, `runContinuous` creates an `IncrementalExecution` (that becomes the `lastExecution`) that is immediately executed (i.e. the entire query execution pipeline is executed up to and including `executedPlan`).

`runContinuous` sets the following local properties:

- `__is_continuous_processing` as `true`
- `__continuous_start_epoch` as the `currentBatchId`
- `__epoch_coordinator_id` as the `currentEpochCoordinatorId`, i.e. `runId` followed by `--` with a random UUID

- `__continuous_epoch_interval` as the interval of the `ContinuousTrigger`

`runContinuous` uses the `EpochCoordinatorRef` helper to create a remote reference to the `EpochCoordinator` RPC endpoint (with the `StreamWriter`, the `ContinuousReader`, the `currentEpochCoordinatorId`, and the `currentBatchId`).

**Note**

The `EpochCoordinator` RPC endpoint runs on the driver as the single point to coordinate epochs across partition tasks.

`runContinuous` creates a daemon `epoch update thread` and starts it immediately.

In `runContinuous` time-tracking section, `runContinuous` requests the physical query plan (of the `IncrementalExecution`) to execute (that simply requests the physical operator to `doExecute` and generate an `RDD[InternalRow]` ).

**Note**

`runContinuous` is used exclusively when `ContinuousExecution` is requested to run an activated streaming query.

## Epoch Update Thread

`runContinuous` creates an `epoch update thread` that...FIXME

## Getting Start Offsets From Checkpoint — `getStartOffsets` Internal Method

```
getStartOffsets(sparkSessionToRunBatches: SparkSession): OffsetSeq
```

`getStartOffsets` ...FIXME

**Note**

`getStartOffsets` is used exclusively when `ContinuousExecution` is requested to run a streaming query in continuous mode.

## Committing Epoch — `commit` Method

```
commit(epoch: Long): Unit
```

In essence, `commit` adds the given epoch to `commit log` and the `committedOffsets`, and requests the `ContinuousReader` to commit the corresponding offset. In the end, `commit` removes old log entries from the `offset` and `commit` logs (to keep `spark.sql.streaming.minBatchesToRetain` entries only).

Internally, `commit recordTriggerOffsets` (with the from and to offsets as the `committedOffsets` and `availableOffsets`, respectively).

At this point, `commit` may simply return when the `stream execution thread` is no longer alive (died).

`commit` requests the `commit log` to store a metadata for the epoch.

`commit` requests the single `ContinuousReader` to `deserialize the offset` for the epoch (from the `offset write-ahead log`).

`commit` adds the single `ContinuousReader` and the offset (for the epoch) to the `committedOffsets` registry.

`commit` requests the single `ContinuousReader` to `commit the offset`.

`commit` requests the `offset` and `commit` logs to `remove log entries` to keep `spark.sql.streaming.minBatchesToRetain` only.

`commit` then acquires the `awaitProgressLock`, wakes up all threads waiting for the `awaitProgressLockCondition` and in the end releases the `awaitProgressLock`.

Note	<code>commit</code> supports only one continuous source (registered in the <code>continuousSources</code> internal registry).
------	---

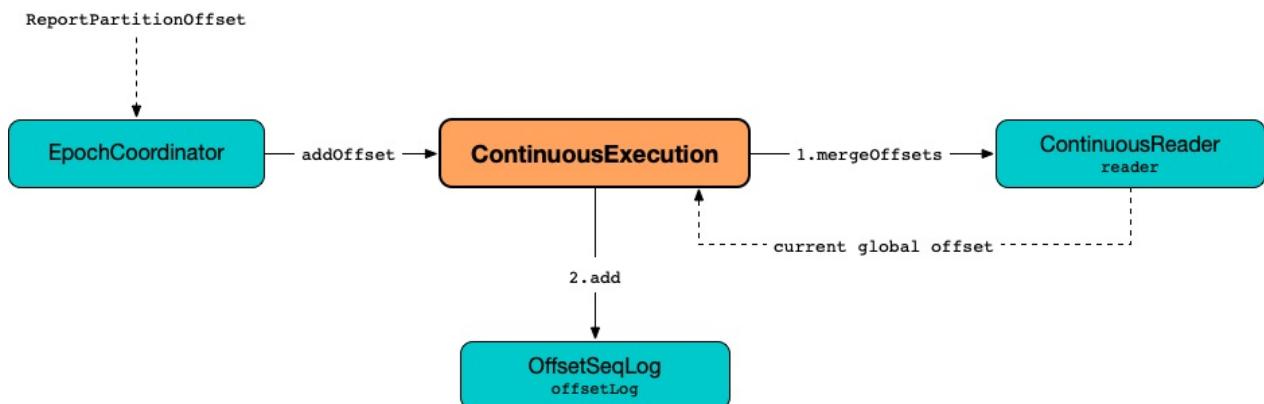
`commit` asserts that the given epoch is available in the `offsetLog` internal registry (i.e. the offset for the given epoch has been reported before).

Note	<code>commit</code> is used exclusively when <code>EpochCoordinator</code> is requested to <code>commitEpoch</code> .
------	---

## addOffset Method

```
addOffset(
  epoch: Long,
  reader: ContinuousReader,
  partitionOffsets: Seq[PartitionOffset]): Unit
```

In essence, `addOffset` requests the given `ContinuousReader` to `mergeOffsets` (with the given `PartitionOffsets`) and then requests the `OffsetSeqLog` to `register the offset with the given epoch`.

Figure 1. `ContinuousExecution.addOffset`

Internally, `addOffset` requests the given `ContinuousReader` to `mergeOffsets` (with the given `PartitionOffsets`) and to get the current "global" offset back.

`addOffset` then requests the `OffsetSeqLog` to `add` the current "global" offset for the given epoch .

`addOffset` requests the `OffsetSeqLog` for the offset at the previous epoch.

If the offsets at the current and previous epochs are the same, `addOffset` turns the `noNewData` internal flag on.

`addOffset` then acquires the `awaitProgressLock`, wakes up all threads waiting for the `awaitProgressLockCondition` and in the end releases the `awaitProgressLock`.

Note	<code>addOffset</code> supports exactly one continuous source.
------	--

Note	<code>addOffset</code> is used exclusively when <code>EpochCoordinator</code> is requested to handle a <code>ReportPartitionOffset</code> message.
------	--

## Analyzed Logical Plan of Streaming Query — `logicalPlan` Property

logicalPlan: LogicalPlan
--------------------------

Note	<code>logicalPlan</code> is part of <code>StreamExecution Contract</code> that is the analyzed logical plan of the streaming query.
------	---

`logicalPlan` resolves `StreamingRelationV2` leaf logical operators (with a `ContinuousReadSupport` source) to `ContinuousExecutionRelation` leaf logical operators.

Internally, `logicalPlan` transforms the `analyzed logical plan` as follows:

1. For every `StreamingRelationV2` leaf logical operator with a `ContinuousReadSupport` source, `logicalPlan` looks it up for the corresponding `ContinuousExecutionRelation` (if available in the internal lookup registry) or creates a `ContinuousExecutionRelation` (with the `ContinuousReadSupport` source, the options and the output attributes of the `StreamingRelationV2` operator)
2. For any other `streamingRelationV2`, `logicalPlan` throws an `UnsupportedOperationException`:

Data source [name] does not support continuous processing.

## Creating ContinuousExecution Instance

`ContinuousExecution` takes the following when created:

- `SparkSession`
- The name of the structured query
- Path to the checkpoint directory (aka *metadata directory*)
- Analyzed logical query plan (`LogicalPlan`)
- `StreamWriterSupport`
- `Trigger`
- `Clock`
- `Output mode`
- Options (`Map[String, String]`)
- `deleteCheckpointOnStop` flag to control whether to delete the checkpoint directory on stop

`ContinuousExecution` initializes the `internal properties`.

## Stopping Stream Processing (Execution of Streaming Query) — `stop` Method

`stop(): Unit`

**Note** `stop` is part of the `StreamingQuery Contract` to stop a streaming query.

`stop` transitions the streaming query to `TERMINATED` state.

If the `queryExecutionThread` is alive (i.e. it has been started and has not yet died), `stop` interrupts it and waits for this thread to die.

In the end, `stop` prints out the following INFO message to the logs:

```
Query [prettyIdString] was stopped
```

Note	<code>prettyIdString</code> is in the format of <code>queryName [id = [id], runId = [runId]]</code> .
------	---

## awaitEpoch Internal Method

```
awaitEpoch(epoch: Long): Unit
```

`awaitEpoch ...FIXME`

Note	<code>awaitEpoch</code> seems to be used exclusively in tests.
------	--

## Internal Properties

Name	Description		
continuousSources	<p>continuousSources: Seq[ContinuousReader]</p> <p>Registry of <a href="#">ContinuousReaders</a> (in the <a href="#">analyzed logical plan of the streaming query</a>)</p> <p>As asserted in <a href="#">commit</a> and <a href="#">addOffset</a> there could only be exactly one <code>ContinuousReaders</code> registered.</p> <p>Used when <code>ContinuousExecution</code> is requested to <a href="#">commit</a>, <a href="#">getStartOffsets</a>, and <a href="#">runContinuous</a></p> <p>Use <a href="#">sources</a> to access the current value</p>		
currentEpochCoordinatorId	<p>FIXME</p> <p>Used when...FIXME</p>		
triggerExecutor	<p>TriggerExecutor for the Trigger:</p> <ul style="list-style-type: none"> <li>• <code>ProcessingTimeExecutor</code> for <a href="#">ContinuousTrigger</a></li> </ul> <p>Used when...FIXME</p> <table border="1"> <tr> <td>Note</td> <td>StreamExecution throws an <code>IllegalStateException</code> when the Trigger is not a <a href="#">ContinuousTrigger</a>.</td> </tr> </table>	Note	StreamExecution throws an <code>IllegalStateException</code> when the Trigger is not a <a href="#">ContinuousTrigger</a> .
Note	StreamExecution throws an <code>IllegalStateException</code> when the Trigger is not a <a href="#">ContinuousTrigger</a> .		

# ContinuousReadSupport Contract — Data Sources with Continuous Readers

`ContinuousReadSupport` is the extension of the `DataSourceV2` for data sources with a `ContinuousReader` for Continuous Stream Processing.

`ContinuousReadSupport` defines a single `createContinuousReader` method to create a `ContinuousReader`.

```
ContinuousReader createContinuousReader(
    Optional<StructType> schema,
    String checkpointLocation,
    DataSourceOptions options)
```

`createContinuousReader` is used when:

- `ContinuousExecution` is requested to run a streaming query (and finds `ContinuousExecutionRelations` in the analyzed logical plan)
- `DataStreamReader` is requested to create a streaming query for a `ContinuousReadSupport` data source

Table 1. ContinuousReadSupports

ContinuousReadSupport	Description
<code>ContinuousMemoryStream</code>	Data source provider for <code>memory</code> format
<code>KafkaSourceProvider</code>	Data source provider for <code>kafka</code> format
<code>RateStreamProvider</code>	Data source provider for <code>rate</code> format
<code>TextSocketSourceProvider</code>	Data source provider for <code>socket</code> format

# ContinuousReader Contract — Data Source Readers in Continuous Stream Processing

`ContinuousReader` is the [extension](#) of Spark SQL's `DataSourceReader` (and `BaseStreamingSource`) contracts for [data source readers](#) in [Continuous Stream Processing](#).

`ContinuousReader` is part of the novel Data Source API V2 in Spark SQL.

Tip	Read up on <a href="#">Data Source API V2</a> in <a href="#">The Internals of Spark SQL</a> book.
-----	---

Table 1. ContinuousReader Contract

Method	Description		
commit	<pre>void commit(offset end)</pre> <p>Commits the specified offset Used exclusively when <code>continuousExecution</code> is requested to <a href="#">commit an epoch</a></p>		
deserializeOffset	<pre>Offset deserializeOffset(String json)</pre> <p>Deserializes an offset from JSON representation Used when <code>ContinuousExecution</code> is requested to <a href="#">run a streaming query</a> and <a href="#">commit an epoch</a></p>		
getStartOffset	<pre>Offset getStartOffset()</pre> <table border="1"> <tr> <td>Note</td> <td>Used exclusively in tests.</td> </tr> </table>	Note	Used exclusively in tests.
Note	Used exclusively in tests.		
mergeOffsets	<pre>Offset mergeOffsets(PartitionOffset[] offsets)</pre> <p>Used exclusively when <code>continuousExecution</code> is requested to <a href="#">addOffset</a></p>		
needsReconfiguration	<pre>boolean needsReconfiguration()</pre> <p>Indicates that the reader needs reconfiguration (e.g. to generate new input partitions) Used exclusively when <code>continuousExecution</code> is requested to <a href="#">run a streaming query in continuous mode</a></p>		
setStartOffset	<pre>void setStartOffset(Optional&lt;Offset&gt; start)</pre> <p>Used exclusively when <code>continuousExecution</code> is requested to <a href="#">run the streaming query in continuous mode</a>.</p>		

Table 2. ContinuousReaders

<b>ContinuousReader</b>	<b>Description</b>
<a href="#">ContinuousMemoryStream</a>	
<a href="#">KafkaContinuousReader</a>	
<a href="#">RateStreamContinuousReader</a>	
<a href="#">TextSocketContinuousReader</a>	

# RateStreamContinuousReader

RateStreamContinuousReader is a [ContinuousReader](#) that...FIXME

# EpochCoordinator RPC Endpoint — Coordinating Epochs and Offsets Across Partition Tasks

`EpochCoordinator` is a `ThreadSafeRpcEndpoint` that tracks offsets and epochs (*coordinates epochs*) by handling [messages](#) (in [fire-and-forget one-way](#) and [request-response two-way](#) modes) from...FIXME

`EpochCoordinator` is created (using `create` factory method) when `ContinuousExecution` is requested to run a streaming query in continuous mode.

Table 1. EpochCoordinator RPC Endpoint's Messages

Message	Description
CommitPartitionEpoch <ul style="list-style-type: none"> <li>• Partition ID</li> <li>• Epoch</li> <li>• DataSource API V2's <code>WriterCommitMessage</code></li> </ul>	Sent out (in one-way asynchronous mode) exclusively when <code>ContinuousWriteRDD</code> is requested to <a href="#">compute a partition</a> (after all rows were written down to a streaming sink)
GetCurrentEpoch	Sent out (in request-response synchronous mode) exclusively when <code>EpochMarkerGenerator</code> thread is requested to <a href="#">run</a>
IncrementAndGetEpoch	Sent out (in request-response synchronous mode) exclusively when <code>ContinuousExecution</code> is requested to <a href="#">run a streaming query in continuous mode</a> (and start a separate epoch update thread)
ReportPartitionOffset <ul style="list-style-type: none"> <li>• Partition ID</li> <li>• Epoch</li> <li>• <a href="#">PartitionOffset</a></li> </ul>	Sent out (in one-way asynchronous mode) exclusively when <code>ContinuousQueuedDataReader</code> is requested for the <a href="#">next row</a> to be read in the current epoch, and the epoch is done
SetReaderPartitions <ul style="list-style-type: none"> <li>• Number of partitions</li> </ul>	Sent out (in request-response synchronous mode) exclusively when <code>DataSourceV2ScanExec</code> leaf physical operator is requested for the input RDDs (for a <a href="#">ContinuousReader</a> and is about to create a <a href="#">ContinuousDataSourceRDD</a> )  The <a href="#">number of partitions</a> is exactly the number of <code>InputPartitions</code> from the <code>ContinuousReader</code> .
SetWriterPartitions <ul style="list-style-type: none"> <li>• Number of partitions</li> </ul>	Sent out (in request-response synchronous mode) exclusively when <code>WriteToContinuousDataSourceExec</code> leaf physical operator is requested to <a href="#">execute</a> and <a href="#">generate a recipe for a distributed computation</a> (as an <code>RDD[InternalRow]</code> ) (and requests a <a href="#">ContinuousWriteRDD</a> to collect that simply never finishes...and that's the <i>trick</i> of continuous mode)
StopContinuousExecutionWrites	Sent out (in request-response synchronous mode) exclusively when <code>ContinuousExecution</code> is requested to <a href="#">run a streaming query in continuous mode</a> (and it finishes successfully or not)

Enable `ALL` logging level for `org.apache.spark.sql.execution.streaming.continuous.EpochCoordinatorRef` logger to see what happens inside.

Add the following line to `conf/log4j.properties`:

Tip

```
log4j.logger.org.apache.spark.sql.execution.streaming.continuous.EpochCoordinatorRef=ALL
```

Refer to [Logging](#).

## Receiving Messages (Fire-And-Forget One-Way Mode)

### — `receive` Method

```
receive: PartialFunction[Any, Unit]
```

Note

`receive` is part of the `RpcEndpoint` Contract in Apache Spark to receive messages in fire-and-forget one-way mode.

`receive` handles the following messages:

- [CommitPartitionEpoch](#)
- [ReportPartitionOffset](#)

With the `queryWritesStopped` turned on, `receive` simply *swallows* messages and does nothing.

## Receiving Messages (Request-Response Two-Way Mode)

### — `receiveAndReply` Method

```
receiveAndReply(context: RpcCallContext): PartialFunction[Any, Unit]
```

Note

`receiveAndReply` is part of the `RpcEndpoint` Contract in Apache Spark to receive and reply to messages in request-response two-way mode.

`receiveAndReply` handles the following messages:

- [GetCurrentEpoch](#)
- [IncrementAndGetEpoch](#)
- [SetReaderPartitions](#)

- [SetWriterPartitions](#)
- [StopContinuousExecutionWrites](#)

## **resolveCommitsAtEpoch Internal Method**

```
resolveCommitsAtEpoch(epoch: Long): Unit
```

resolveCommitsAtEpoch ...FIXME

Note

`resolveCommitsAtEpoch` is used exclusively when `EpochCoordinator` is requested to handle [CommitPartitionEpoch](#) and [ReportPartitionOffset](#) messages.

## **commitEpoch Internal Method**

```
commitEpoch(
  epoch: Long,
  messages: Iterable[WriterCommitMessage]): Unit
```

commitEpoch ...FIXME

Note

`commitEpoch` is used exclusively when `EpochCoordinator` is requested to [resolveCommitsAtEpoch](#).

## **Creating EpochCoordinator Instance**

`EpochCoordinator` takes the following to be created:

- [StreamWriter](#)
- [ContinuousReader](#)
- [ContinuousExecution](#)
- Start epoch
- [SparkSession](#)
- [RpcEnv](#)

`EpochCoordinator` initializes the [internal properties](#).

## Registering EpochCoordinator RPC Endpoint— `create` Factory Method

```
create(
    writer: StreamWriter,
    reader: ContinuousReader,
    query: ContinuousExecution,
    epochCoordinatorId: String,
    startEpoch: Long,
    session: SparkSession,
    env: SparkEnv): RpcEndpointRef
```

`create` simply [creates a new EpochCoordinator](#) and requests the `RpcEnv` to register a RPC endpoint as **EpochCoordinator-[id]** (where `id` is the given `epochCoordinatorId` ).

`create` prints out the following INFO message to the logs:

```
Registered EpochCoordinator endpoint
```

Note	<code>create</code> is used exclusively when <code>ContinuousExecution</code> is requested to <a href="#">run a streaming query in continuous mode</a> .
------	--

## Internal Properties

Name	Description
queryWritesStopped	<p>Flag that indicates whether to drop messages (<code>true</code>) or not (<code>false</code>) when requested to <a href="#">handle one synchronously</a></p> <p>Default: <code>false</code></p> <p>Turned on (<code>true</code>) when requested to <a href="#">handle a synchronous StopContinuousExecutionWrites message</a></p>

# EpochCoordinatorRef

`EpochCoordinatorRef` is...FIXME

## Creating Remote Reference to EpochCoordinator RPC Endpoint— `create` Factory Method

```
create(
    writer: StreamWriter,
    reader: ContinuousReader,
    query: ContinuousExecution,
    epochCoordinatorId: String,
    startEpoch: Long,
    session: SparkSession,
    env: SparkEnv): RpcEndpointRef
```

`create` ...FIXME

Note

`create` is used exclusively when `ContinuousExecution` is requested to run a streaming query in continuous mode.

## Getting Remote Reference to EpochCoordinator RPC Endpoint— `get` Factory Method

```
get(id: String, env: SparkEnv): RpcEndpointRef
```

`get` ...FIXME

Note

`get` is used when:

- `DataSourceV2ScanExec` leaf physical operator is requested for the input RDDs (and creates a `ContinuousDataSourceRDD` for a `ContinuousReader`)
- `ContinuousQueuedDataReader` is created (and initializes the `epochCoordEndpoint`)
- `EpochMarkerGenerator` is created (and initializes the `epochCoordEndpoint`)
- `ContinuousWriteRDD` is requested to compute a partition
- `WriteToContinuousDataSourceExec` is requested to execute and generate a recipe for a distributed computation (as an `RDD[InternalRow]`)



# EpochTracker

EpochTracker is...FIXME

## Current Epoch — `getCurrentEpoch` Method

```
getCurrentEpoch: Option[Long]
```

getCurrentEpoch ...FIXME

Note

`getCurrentEpoch` is used when...FIXME

## Advancing (Incrementing) Epoch — `incrementCurrentEpoch` Method

```
incrementCurrentEpoch(): Unit
```

incrementCurrentEpoch ...FIXME

Note

`incrementCurrentEpoch` is used when...FIXME

# ContinuousQueuedDataReader

`ContinuousQueuedDataReader` is created exclusively when `ContinuousDataSourceRDD` is requested to compute a partition.

`ContinuousQueuedDataReader` uses two types of continuous records:

- `EpochMarker`
- `ContinuousRow` (with the `InternalRow` at `PartitionOffset`)

## Fetching Next Row — `next` Method

```
next(): InternalRow
```

`next` ...FIXME

Note	<code>next</code> is used when...FIXME
------	--

## Closing ContinuousQueuedDataReader — `close` Method

```
close(): Unit
```

Note	<code>close</code> is part of the <a href="#">java.io.Closeable</a> to close this stream and release any system resources associated with it.
------	---

`close` ...FIXME

## Creating ContinuousQueuedDataReader Instance

`ContinuousQueuedDataReader` takes the following to be created:

- `ContinuousDataSourceRDDPartition`
- `TaskContext`
- Size of the [data queue](#)
- `epochPollIntervalMs`

`ContinuousQueuedDataReader` initializes the [internal properties](#).

## Internal Properties

Name	Description
coordinatorId	<b>Epoch Coordinator Identifier</b> Used when...FIXME
currentOffset	PartitionOffset Used when...FIXME
dataReaderThread	<a href="#">DataReaderThread</a> daemon thread that is created and started immediately when <code>ContinuousQueuedDataReader</code> is created Used when...FIXME
epochCoordEndpoint	<code>RpcEndpointRef</code> of the <a href="#">EpochCoordinator</a> per <code>coordinatorId</code> Used when...FIXME
epochMarkerExecutor	<a href="#">java.util.concurrent.ScheduledExecutorService</a> Used when...FIXME
epochMarkerGenerator	<a href="#">EpochMarkerGenerator</a> Used when...FIXME
reader	<a href="#">InputPartitionReader</a> Used when...FIXME
queue	<a href="#">java.util.concurrent.ArrayBlockingQueue</a> of <a href="#">ContinuousRecords</a> (of the given <code>data size</code> ) Used when...FIXME

# DataReaderThread

DataReaderThread is...FIXME

# EpochMarkerGenerator Thread

EpochMarkerGenerator is...FIXME

## run Method

run(): Unit

Note

run is part of the [java.lang.Runnable](#) Contract to be executed upon starting a thread.

run ...FIXME

# PartitionOffset

PartitionOffset is...FIXME

# ContinuousExecutionRelation Leaf Logical Operator

`ContinuousExecutionRelation` is a `MultiInstanceRelation` leaf logical operator.

Tip	Read up on <a href="#">Leaf Logical Operators</a> in <a href="#">The Internals of Spark SQL</a> book.
-----	---

`ContinuousExecutionRelation` is [created](#) (to represent `StreamingRelationV2` with `ContinuousReadSupport` data source) when `ContinuousExecution` is [created](#) (and requested for the [logical plan](#)).

`ContinuousExecutionRelation` takes the following to be created:

- [ContinuousReadSupport](#) source
- Options ( `Map[String, String]` )
- Output attributes ( `seq[Attribute]` )
- `SparkSession`

# WriteToContinuousDataSource Unary Logical Operator

`WriteToContinuousDataSource` is a unary logical operator (`LogicalPlan`) that is created exclusively when `ContinuousExecution` is requested to run a streaming query in continuous mode (to create an `IncrementalExecution`).

`WriteToContinuousDataSource` is planned (*translated*) to a `WriteToContinuousDataSourceExec` unary physical operator (when `DataSourceV2Strategy` execution planning strategy is requested to plan a logical query).

Tip

Read up on [DataSourceV2Strategy Execution Planning Strategy](#) in [The Internals of Spark SQL book](#).

`WriteToContinuousDataSource` takes the following to be created:

- [StreamWriter](#)
- Child logical operator (`LogicalPlan`)

`WriteToContinuousDataSource` uses empty output schema (which is exactly to say that no output is expected whatsoever).

# WriteToContinuousDataSourceExec Unary Physical Operator

`WriteToContinuousDataSourceExec` is a unary physical operator that [creates a ContinuousWriteRDD for continuous write](#).

Note

A unary physical operator (`UnaryExecNode`) is a physical operator with a single [child](#) physical operator.

Read up on [UnaryExecNode](#) (and physical operators in general) in [The Internals of Spark SQL](#) book.

`WriteToContinuousDataSourceExec` is [created](#) exclusively when `DataSourceV2Strategy` execution planning strategy is requested to plan a [WriteToContinuousDataSource](#) unary logical operator.

Tip

Read up on [DataSourceV2Strategy Execution Planning Strategy](#) in [The Internals of Spark SQL](#) book.

`WriteToContinuousDataSourceExec` takes the following to be created:

- [StreamWriter](#)
- Child physical operator (`SparkPlan`)

`WriteToContinuousDataSourceExec` uses empty output schema (which is exactly to say that no output is expected whatsoever).

Tip

Enable `ALL` logging level for

```
org.apache.spark.sql.execution.streaming.continuous.WriteToContinuousDataSourceExec
```

happens inside.

Add the following line to `conf/log4j.properties` :

```
log4j.logger.org.apache.spark.sql.execution.streaming.continuous.WriteToContinuousDa
```

Refer to [Logging](#).

## Executing Physical Operator (Generating RDD[InternalRow]) — `doExecute` Method

```
doExecute(): RDD[InternalRow]
```

**Note** `doExecute` is part of `SparkPlan` Contract to generate the runtime representation of an physical operator as a distributed computation over internal binary rows on Apache Spark (i.e. `RDD[InternalRow]` ).

`doExecute` requests the [StreamWriter](#) to create a `DataWriterFactory`.

`doExecute` then requests the [child physical operator](#) to execute (that gives a `RDD[InternalRow]` ) and uses the `RDD[InternalRow]` and the `DataWriterFactory` to create a [ContinuousWriteRDD](#).

`doExecute` prints out the following INFO message to the logs:

```
Start processing data source writer: [writer]. The input RDD has [partitions] partitions.
```

`doExecute` requests the `EpochCoordinatorRef` helper for a [remote reference to the EpochCoordinator RPC endpoint](#) (using the `_epoch_coordinator_id` local property).

**Note** The [EpochCoordinator RPC endpoint](#) runs on the driver as the single point to coordinate epochs across partition tasks.

`doExecute` requests the EpochCoordinator RPC endpoint reference to send out a [SetWriterPartitions](#) message synchronously.

In the end, `doExecute` requests the `ContinuousWriteRDD` to collect (which simply runs a Spark job on all partitions in an RDD and returns the results in an array).

**Note** Requesting the `ContinuousWriteRDD` to collect is how a Spark job is ran that in turn runs tasks (one per partition) that are described by the `ContinuousWriteRDD.compute` method. Since executing `collect` is meant to run a Spark job (with tasks on executors), it's in the discretion of the tasks themselves to decide when to finish (so if they want to run indefinitely, so be it). *What a clever trick!*

# ContinuousWriteRDD — RDD of WriteToContinuousDataSourceExec Unary Physical Operator

`ContinuousWriteRDD` is a specialized `RDD` (`RDD[Unit]`) that is used exclusively as the underlying `RDD` of `writeToContinuousDataSourceExec` unary physical operator to [write records continuously](#).

`ContinuousWriteRDD` is [created](#) exclusively when `writeToContinuousDataSourceExec` unary physical operator is requested to [execute](#) and generate a recipe for a distributed computation (as an `RDD[InternalRow]`).

`ContinuousWriteRDD` uses the [parent RDD](#) for the partitions and the partitioner.

`ContinuousWriteRDD` takes the following to be created:

- Parent `RDD` (`RDD[InternalRow]`)
- Write task (`DataWriterFactory[InternalRow]`)

## Computing Partition — `compute` Method

```
compute(  
    split: Partition,  
    context: TaskContext): Iterator[Unit]
```

Note	<code>compute</code> is part of the <code>RDD</code> Contract to compute a partition.
------	---

`compute` requests the `EpochCoordinatorRef` helper for a [remote reference to the EpochCoordinator RPC endpoint](#) (using the `__epoch_coordinator_id` local property).

Note	The <a href="#">EpochCoordinator RPC endpoint</a> runs on the driver as the single point to coordinate epochs across partition tasks.
------	---

`compute` uses the `EpochTracker` helper to [initializeCurrentEpoch](#) (using the `__continuous_start_epoch` local property).

`compute` then executes the following steps (in a loop) until the task (as the given `TaskContext`) is killed or completed.

`compute` requests the [parent RDD](#) to compute the given partition (that gives an `Iterator[InternalRow]`).

`compute` requests the [DataWriterFactory](#) to create a `DataWriter` (for the partition and the task attempt IDs from the given `TaskContext` and the [current epoch](#) from the `EpochTracker` helper) and requests it to write all records (from the `Iterator[InternalRow]` ).

`compute` prints out the following INFO message to the logs:

```
Writer for partition [partitionId] in epoch [epoch] is committing.
```

`compute` requests the `DataWriter` to commit (that gives a `WriterCommitMessage` ).

`compute` requests the EpochCoordinator RPC endpoint reference to send out a [CommitPartitionEpoch](#) message (with the `WriterCommitMessage` ).

`compute` prints out the following INFO message to the logs:

```
Writer for partition [partitionId] in epoch [epoch] is committed.
```

In the end (of the loop), `compute` uses the `EpochTracker` helper to [incrementCurrentEpoch](#).

In case of an error, `compute` prints out the following ERROR message to the logs and requests the `DataWriter` to abort.

```
Writer for partition [partitionId] is aborting.
```

In the end, `compute` prints out the following ERROR message to the logs:

```
Writer for partition [partitionId] aborted.
```

# ContinuousDataSourceRDD — Input RDD of DataSourceV2ScanExec Physical Operator with ContinuousReader

`ContinuousDataSourceRDD` is a specialized `RDD` (`RDD[InternalRow]`) that is used exclusively for the only input RDD (with the input rows) of `DataSourceV2ScanExec` leaf physical operator with a [ContinuousReader](#).

`ContinuousDataSourceRDD` is [created](#) exclusively when `DataSourceV2ScanExec` leaf physical operator is requested for the input RDDs (which there is only one actually).

`ContinuousDataSourceRDD` uses [spark.sql.streaming.continuous.executorQueueSize](#) configuration property for the size of the data queue.

`ContinuousDataSourceRDD` uses [spark.sql.streaming.continuous.executorPollIntervalMs](#) configuration property for the epochPollIntervalMs.

`ContinuousDataSourceRDD` takes the following to be created:

- `SparkContext`
- Size of the data queue
- `epochPollIntervalMs`
- `InputPartition[InternalRow] S`

`ContinuousDataSourceRDD` uses `InputPartition` (of a `ContinuousDataSourceRDDPartition`) for preferred host locations (where the input partition reader can run faster).

## Computing Partition — `compute` Method

```
compute(
  split: Partition,
  context: TaskContext): Iterator[InternalRow]
```

Note

`compute` is part of the RDD Contract to compute a given partition.

`compute` ...FIXME

## getPartitions Method

```
getPartitions: Array[Partition]
```

<b>Note</b>	getPartitions is part of the <code>RDD</code> Contract to specify the partitions to <a href="#">compute</a> .
-------------	---

```
getPartitions ...FIXME
```

# StreamExecution — Base of Stream Execution Engines

`StreamExecution` is the [base](#) of [stream execution engines](#) (aka *streaming query processing engines*) that can [run](#) a [structured query](#) (on a [stream execution thread](#)).

Note	<b>Continuous query, streaming query, continuous Dataset, streaming Dataset</b> are all considered high-level synonyms for an executable entity that stream execution engines run using the <a href="#">analyzed logical plan</a> internally.
------	---

Table 1. StreamExecution Contract (Abstract Methods Only)

Property	Description
<code>logicalPlan</code>	<pre>logicalPlan: LogicalPlan</pre> <p>Analyzed logical plan of the streaming query to execute Used when <code>StreamExecution</code> is requested to <a href="#">run stream processing</a></p> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 10px;"> <p><b>Note</b> <code>logicalPlan</code> is part of <a href="#">ProgressReporter Contract</a> and the only purpose of the <code>logicalPlan</code> property is to change the access level from <code>protected</code> to <code>public</code>.</p> </div>
<code>runActivatedStream</code>	<pre>runActivatedStream(     sparkSessionForStream: SparkSession): Unit</pre> <p>Executes (<i>runs</i>) the activated <a href="#">streaming query</a> Used exclusively when <code>StreamExecution</code> is requested to <a href="#">run the streaming query</a> (when transitioning from <code>INITIALIZING</code> to <code>ACTIVE</code> state)</p>

## Streaming Query and Stream Execution Engine

```
import org.apache.spark.sql.streaming.StreamingQuery
assert(sq.isInstanceOf[StreamingQuery])

import org.apache.spark.sql.execution.streaming.StreamingQueryWrapper
val se = sq.asInstanceOf[StreamingQueryWrapper].streamingQuery

scala> :type se
org.apache.spark.sql.execution.streaming.StreamExecution
```

`StreamExecution` uses the `spark.sql.streaming.minBatchesToRetain` configuration property to allow the `StreamExecutions` to discard old log entries (from the `offset` and `commit` logs).

Table 2. StreamExecutions

StreamExecution	Description
ContinuousExecution	Used in <a href="#">Continuous Stream Processing</a>
MicroBatchExecution	Used in <a href="#">Micro-Batch Stream Processing</a>

Note	<code>StreamExecution</code> does not support adaptive query execution and cost-based optimizer (and turns them off when requested to <a href="#">run stream processing</a> ).
------	--

`StreamExecution` is the **execution environment** of a single streaming query (aka *streaming Dataset*) that is executed every `trigger` and in the end adds the results to a `sink`.

Note	<code>StreamExecution</code> corresponds to a single streaming query with one or more <a href="#">streaming sources</a> and exactly one <a href="#">streaming sink</a> .
------	--

```

import org.apache.spark.sql.streaming.Trigger
import scala.concurrent.duration._
val q = spark.
  readStream.
  format("rate").
  load.
  writeStream.
  format("console").
  trigger(Trigger.ProcessingTime(10.minutes)).
  start
scala> :type q
org.apache.spark.sql.streaming.StreamingQuery

// Pull out StreamExecution off StreamingQueryWrapper
import org.apache.spark.sql.execution.streaming.{StreamExecution, StreamingQueryWrapper}
}
val se = q.asInstanceOf[StreamingQueryWrapper].streamingQuery
scala> :type se
org.apache.spark.sql.execution.streaming.StreamExecution

```

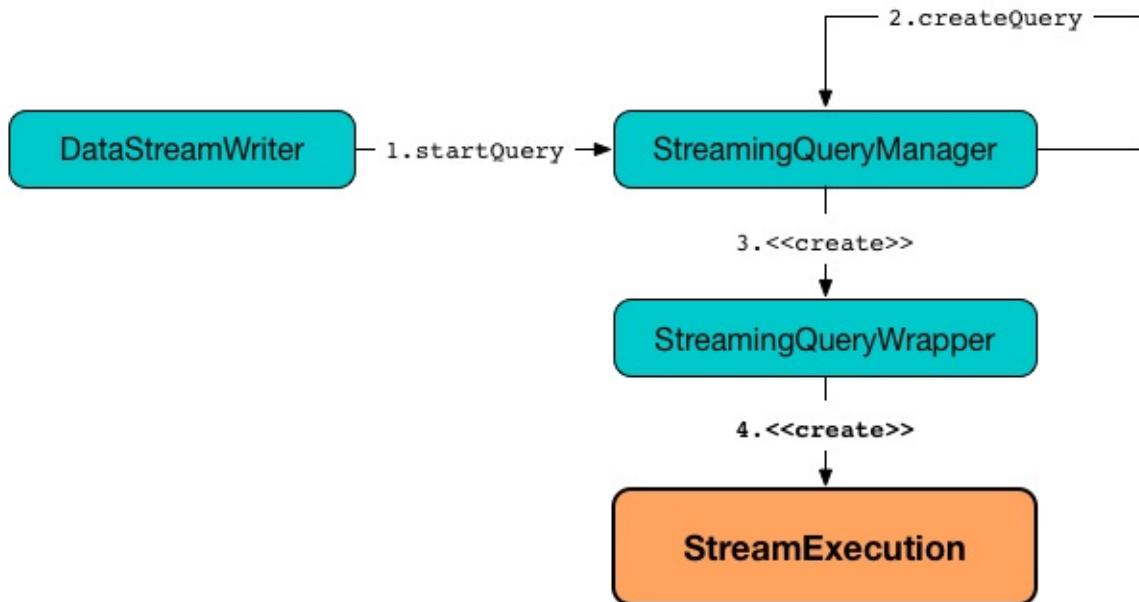


Figure 1. Creating Instance of StreamExecution

**Note**

`DataStreamWriter` describes how the results of executing batches of a streaming query are written to a streaming sink.

When `started`, `StreamExecution` starts a `stream execution thread` that simply `runs stream processing` (and hence the streaming query).

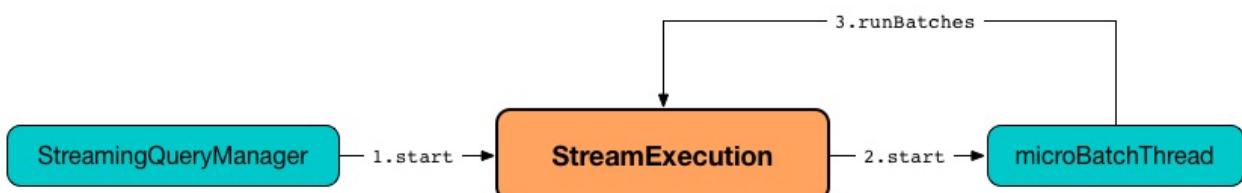


Figure 2. StreamExecution's Starting Streaming Query (on Execution Thread)

`StreamExecution` is a `ProgressReporter` and `reports status of the streaming query` (i.e. when it starts, progresses and terminates) by posting `StreamingQueryListener` events.

```

import org.apache.spark.sql.streaming.Trigger
import scala.concurrent.duration._

val sq = spark
  .readStream
  .text("server-logs")
  .writeStream
  .format("console")
  .queryName("debug")
  .trigger(Trigger.ProcessingTime(20.seconds))
  .start

// Enable the log level to see the INFO and DEBUG messages
// log4j.logger.org.apache.spark.sql.execution.streaming.StreamExecution=DEBUG

17/06/18 21:21:07 INFO StreamExecution: Starting new streaming query.
17/06/18 21:21:07 DEBUG StreamExecution: getOffset took 5 ms
17/06/18 21:21:07 DEBUG StreamExecution: Stream running from {} to {}
17/06/18 21:21:07 DEBUG StreamExecution: triggerExecution took 9 ms
17/06/18 21:21:07 DEBUG StreamExecution: Execution stats: ExecutionStats(Map(),List(),
Map())
17/06/18 21:21:07 INFO StreamExecution: Streaming query made progress: {
  "id" : "8b57b0bd-fc4a-42eb-81a3-777d7ba5e370",
  "runId" : "920b227e-6d02-4a03-a271-c62120258cea",
  "name" : "debug",
  "timestamp" : "2017-06-18T19:21:07.693Z",
  "numInputRows" : 0,
  "processedRowsPerSecond" : 0.0,
  "durationMs" : {
    "getOffset" : 5,
    "triggerExecution" : 9
  },
  "stateOperators" : [ ],
  "sources" : [ {
    "description" : "FileStreamSource[file:/Users/jacek/dev/oss/spark/server-logs]",
    "startOffset" : null,
    "endOffset" : null,
    "numInputRows" : 0,
    "processedRowsPerSecond" : 0.0
  }],
  "sink" : {
    "description" : "org.apache.spark.sql.execution.streaming.ConsoleSink@2460208a"
  }
}
17/06/18 21:21:10 DEBUG StreamExecution: Starting Trigger Calculation
17/06/18 21:21:10 DEBUG StreamExecution: getOffset took 3 ms
17/06/18 21:21:10 DEBUG StreamExecution: triggerExecution took 3 ms
17/06/18 21:21:10 DEBUG StreamExecution: Execution stats: ExecutionStats(Map(),List(),
Map())

```

`StreamExecution` tracks streaming data sources in `uniqueSources` internal registry.

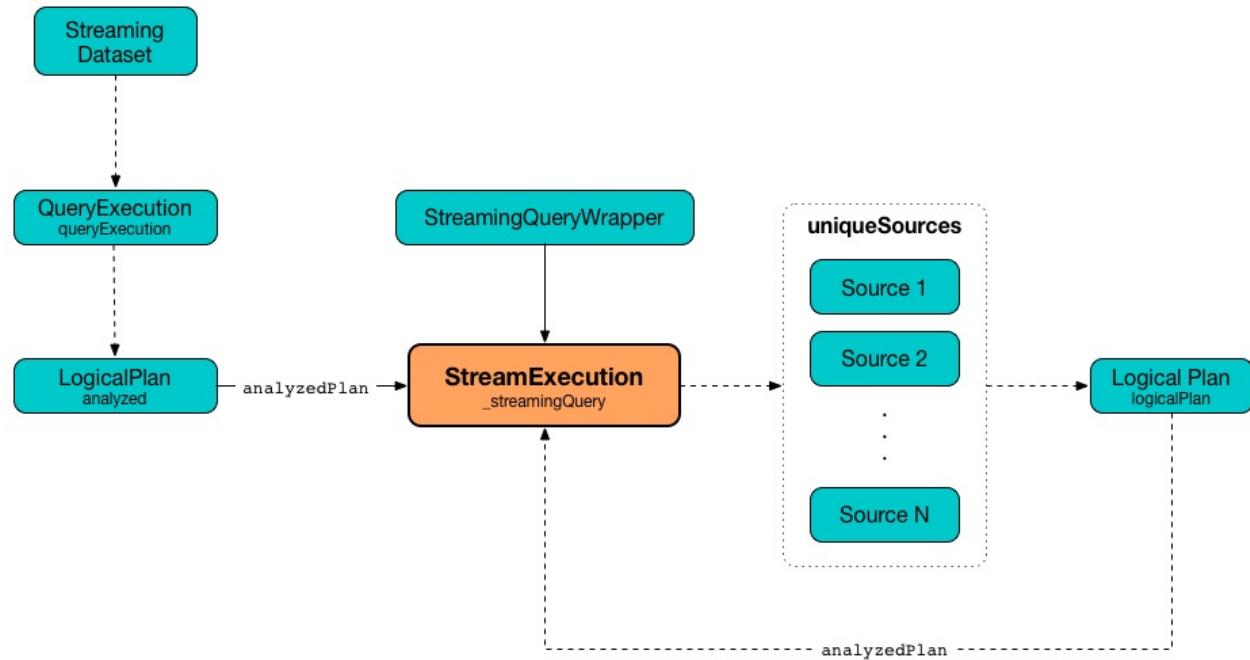


Figure 3. StreamExecution's uniqueSources Registry of Streaming Data Sources

`StreamExecution` collects `durationMs` for the execution units of streaming batches.

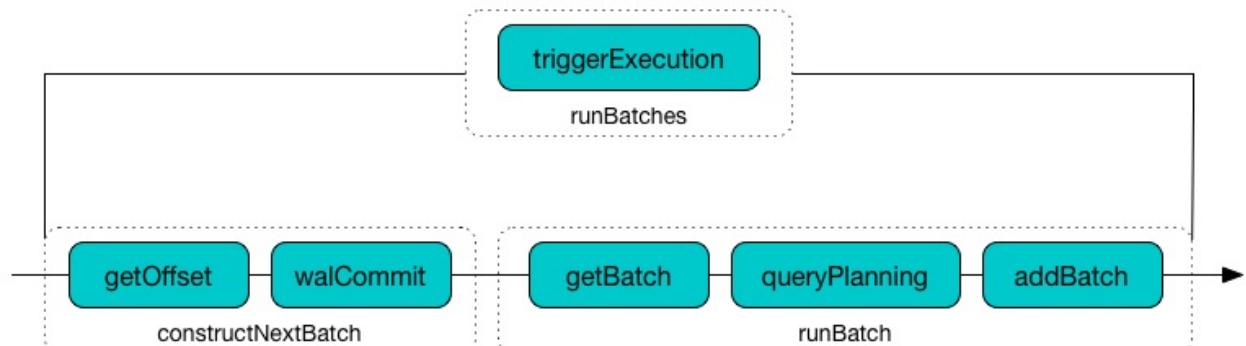


Figure 4. StreamExecution's durationMs

```

scala> :type q
org.apache.spark.sql.streaming.StreamingQuery

scala> println(q.lastProgress)
{
  "id" : "03fc78fc-fe19-408c-a1ae-812d0e28fcee",
  "runId" : "8c247071-afba-40e5-aad2-0e6f45f22488",
  "name" : null,
  "timestamp" : "2017-08-14T20:30:00.004Z",
  "batchId" : 1,
  "numInputRows" : 432,
  "inputRowsPerSecond" : 0.9993568953312452,
  "processedRowsPerSecond" : 1380.1916932907347,
  "durationMs" : {
    "addBatch" : 237,
    "getBatch" : 26,
    "getOffset" : 0,
    "queryPlanning" : 1,
    "triggerExecution" : 313,
    "walCommit" : 45
  },
  "stateOperators" : [ ],
  "sources" : [ {
    "description" : "RateSource[rowsPerSecond=1, rampUpTimeSeconds=0, numPartitions=8]"
  },
  {
    "startOffset" : 0,
    "endOffset" : 432,
    "numInputRows" : 432,
    "inputRowsPerSecond" : 0.9993568953312452,
    "processedRowsPerSecond" : 1380.1916932907347
  } ],
  "sink" : {
    "description" : "ConsoleSink[numRows=20, truncate=true]"
  }
}

```

`StreamExecution` uses [OffsetSeqLog](#) and [BatchCommitLog](#) metadata logs for **write-ahead log** (to record offsets to be processed) and that have already been processed and committed to a streaming sink, respectively.

Tip

Monitor `offsets` and `commits` metadata logs to know the progress of a streaming query.

`StreamExecution` delays polling for new data for 10 milliseconds (when no data was available to process in a batch). Use `spark.sql.streaming.pollingDelay` Spark property to control the delay.

Every `StreamExecution` is uniquely identified by an **ID of the streaming query** (which is the `id` of the [StreamMetadata](#)).

**Note**

Since the [StreamMetadata](#) is persisted (to the `metadata` file in the [checkpoint directory](#)), the streaming query ID "survives" query restarts as long as the checkpoint directory is preserved.

`StreamExecution` is also uniquely identified by a **run ID of the streaming query**. A run ID is a randomly-generated 128-bit universally unique identifier (UUID) that is assigned at the time `StreamExecution` is created.

**Note**

`runId` does not "survive" query restarts and will always be different yet unique (across all active queries).

**Note**

The `name`, `id` and `runId` are all unique across all active queries (in a [StreamingQueryManager](#)). The difference is that:

- `name` is optional and user-defined
- `id` is a UUID that is auto-generated at the time `StreamExecution` is created and persisted to `metadata` checkpoint file
- `runId` is a UUID that is auto-generated every time `StreamExecution` is created

`StreamExecution` uses a [StreamMetadata](#) that is [persisted](#) in the `metadata` file in the [checkpoint directory](#). If the `metadata` file is available it is [read](#) and is the way to recover the [ID](#) of a streaming query when resumed (i.e. restarted after a failure or a planned stop).

`StreamExecution` uses [\\_\\_is\\_continuous\\_processing](#) local property (default: `false`) to differentiate between [ContinuousExecution](#) (`true`) and [MicroBatchExecution](#) (`false`) which is used when `statestoreRDD` is requested to [compute a partition](#) (and [finds a StateStore](#) for a given version).

**Tip**

Enable `ALL` logging level for `org.apache.spark.sql.execution.streaming.StreamExecution` to see what happens inside.

Add the following line to `conf/log4j.properties`:

```
log4j.logger.org.apache.spark.sql.execution.streaming.StreamExecution=ALL
```

Refer to [Logging](#).

## Creating StreamExecution Instance

`StreamExecution` takes the following to be created:

- `SparkSession`
- Name of the streaming query (can also be `null`)
- Path of the checkpoint directory (aka *metadata directory*)
- Streaming query (as an analyzed logical query plan, i.e. `LogicalPlan`)
- **Streaming sink**
- **Trigger**
- `Clock`
- **Output mode**
- `deleteCheckpointOnStop` flag (to control whether to delete the checkpoint directory on stop)

`StreamExecution` initializes the [internal properties](#).

Note	<code>StreamExecution</code> is a Scala abstract class and cannot be <a href="#">created</a> directly. It is created indirectly when the <a href="#">concrete StreamExecutions</a> are.
------	---

## Write-Ahead Log (WAL) of Offsets — `offsetLog` Property

<pre>offsetLog: OffsetSeqLog</pre>
------------------------------------

`offsetLog` is a [Hadoop DFS-based metadata storage](#) (of `OffsetSeqs`) with [offsets metadata directory](#).

`offsetLog` is used as **Write-Ahead Log of Offsets** to [persist offsets](#) of the data about to be processed in every trigger.

Note	<b>Metadata log</b> or <b>metadata checkpoint</b> are synonyms and are often used interchangeably.
------	--

The number of entries in the `offsetSeqLog` is controlled using `spark.sql.streaming.minBatchesToRetain` configuration property (default: `100`). [Stream execution engines](#) discard ([purge](#)) offsets from the `offsets` metadata log when the [current batch ID](#) (in [MicroBatchExecution](#)) or the [epoch committed](#) (in [ContinuousExecution](#)) is above the threshold.

Note	<p><code>offsetLog</code> is used when:</p> <ul style="list-style-type: none"> <li>• <code>ContinuousExecution</code> stream execution engine is requested to commit an epoch, <code>getStartOffsets</code>, and <code>addOffset</code></li> <li>• <code>MicroBatchExecution</code> stream execution engine is requested to populate start offsets and construct (or skip) the next streaming micro-batch</li> </ul>
------	--

## State of Streaming Query (Execution) — `state` Property

```
state: AtomicReference[State]
```

`state` indicates the internal state of execution of the streaming query (as `java.util.concurrent.atomic.AtomicReference`).

Table 3. States

Name	Description
ACTIVE	<code>StreamExecution</code> has been requested to run stream processing (and is about to run the activated streaming query)
INITIALIZING	<code>StreamExecution</code> has been created
TERMINATED	Used to indicate that: <ul style="list-style-type: none"> <li>• <code>MicroBatchExecution</code> has been requested to stop</li> <li>• <code>ContinuousExecution</code> has been requested to stop</li> <li>• <code>StreamExecution</code> has been requested to run stream processing (and has finished running the activated streaming query)</li> </ul>
RECONFIGURING	Used only when <code>ContinuousExecution</code> is requested to run a streaming query in continuous mode (and the <code>ContinuousReader</code> indicated a need for reconfiguration)

## Available Offsets (StreamProgress) — `availableOffsets` Property

```
availableOffsets: StreamProgress
```

`availableOffsets` is a collection of offsets per streaming source to track what data (by offset) is available for processing for every streaming source in the streaming query (and have not yet been committed).

`availableOffsets` works in tandem with the `committedOffsets` internal registry.

`availableOffsets` is empty when `streamExecution` is created (i.e. no offsets are reported for any streaming source in the streaming query).

Note	<p><code>availableOffsets</code> is used when:</p> <ul style="list-style-type: none"> <li>• <code>MicroBatchExecution</code> stream execution engine is requested to resume and fetch the start offsets from checkpoint, check whether new data is available, construct the next streaming micro-batch and run a single streaming micro-batch</li> <li>• <code>ContinuousExecution</code> stream execution engine is requested to commit an epoch</li> <li>• <code>StreamExecution</code> is requested for the internal string representation</li> </ul>
------	--

## Committed Offsets (StreamProgress) — `committedOffsets` Property

`committedOffsets: StreamProgress`

`committedOffsets` is a collection of offsets per streaming source to track what data (by offset) has already been processed and committed (to the sink or state stores) for every streaming source in the streaming query.

`committedOffsets` works in tandem with the `availableOffsets` internal registry.

Note	<p><code>committedOffsets</code> is used when:</p> <ul style="list-style-type: none"> <li>• <code>MicroBatchExecution</code> stream execution engine is requested for the start offsets (from checkpoint), to check whether new data is available and run a single streaming micro-batch</li> <li>• <code>ContinuousExecution</code> stream execution engine is requested for the start offsets (from checkpoint) and to commit an epoch</li> <li>• <code>StreamExecution</code> is requested for the internal string representation</li> </ul>
------	---

## Fully-Qualified (Resolved) Path to Checkpoint Root Directory — `resolvedCheckpointRoot` Property

```
resolvedCheckpointRoot: String
```

`resolvedCheckpointRoot` is a fully-qualified path of the given [checkpoint root directory](#).

The given [checkpoint root directory](#) is defined using **checkpointLocation** option or the `spark.sql.streaming.checkpointLocation` configuration property with `queryName` option.

`checkpointLocation` and `queryName` options are defined when `StreamingQueryManager` is requested to [create a streaming query](#).

`resolvedCheckpointRoot` is used when [creating the path to the checkpoint directory](#) and when `StreamExecution` finishes [running streaming batches](#).

`resolvedCheckpointRoot` is used for the [logicalPlan](#) (while transforming [analyzedPlan](#) and planning `StreamingRelation` logical operators to corresponding `StreamingExecutionRelation` physical operators with the streaming data sources created passing in the path to `sources` directory to store checkpointing metadata).

	You can see <code>resolvedCheckpointRoot</code> in the INFO message when <code>StreamExecution</code> started.
--	--

Tip

Starting [prettyIdString]. Use [resolvedCheckpointRoot] to store the query check

Internally, `resolvedCheckpointRoot` creates a Hadoop `org.apache.hadoop.fs.Path` for [checkpointRoot](#) and makes it qualified.

Note	<code>resolvedCheckpointRoot</code> uses <code>SparkSession</code> to access <code>SessionState</code> for a Hadoop configuration.
------	--

## Offset Commit Log — `commits` Metadata Checkpoint Directory

`StreamExecution` uses **offset commit log** ([CommitLog](#) with `commits` [metadata checkpoint directory](#)) for streaming batches successfully executed (with a single file per batch with a file name being the batch id) or committed epochs.

Note	<b>Metadata log</b> or <b>metadata checkpoint</b> are synonyms and are often used interchangeably.
------	--

`commitLog` is used by the [stream execution engines](#) for the following:

- `MicroBatchExecution` is requested to [run an activated streaming query](#) (that in turn requests to [populate the start offsets](#) at the very beginning of the streaming query execution and later regularly every [single batch](#))
- `ContinuousExecution` is requested to [run an activated streaming query in continuous mode](#) (that in turn requests to [retrieve the start offsets](#) at the very beginning of the streaming query execution and later regularly every [commit](#))

## Last Query Execution Of Streaming Query (`IncrementalExecution`) — `lastExecution` Property

`lastExecution: IncrementalExecution`

Note	<code>lastExecution</code> is part of the <a href="#">ProgressReporter Contract</a> for the <code>QueryExecution</code> of a streaming query.
------	---

`lastExecution` is a [IncrementalExecution](#) (a `QueryExecution` of a streaming query) of the most recent (*last*) execution.

`lastExecution` is created when the [stream execution engines](#) are requested for the following:

- `MicroBatchExecution` is requested to [run a single streaming micro-batch](#) (when in [queryPlanning Phase](#))
- `ContinuousExecution` stream execution engine is requested to [run a streaming query](#) (when in [queryPlanning Phase](#))

`lastExecution` is used when:

- `StreamExecution` is requested to [explain a streaming query](#) (via `explainInternal`)
- `ProgressReporter` is requested to [extractStateOperatorMetrics](#), [extractExecutionStats](#), and [extractSourceToNumInputRows](#)
- `MicroBatchExecution` stream execution engine is requested to [construct or skip the next streaming micro-batch](#) (based on [StateStoreWriters in a streaming query](#)), [run a single streaming micro-batch](#) (when in [addBatch Phase](#) and [updating watermark and committing offsets to offset commit log](#))
- `ContinuousExecution` stream execution engine is requested to [run a streaming query](#) (when in [runContinuous Phase](#))
- For debugging query execution of streaming queries (using `debugCodegen` )

## Explaining Streaming Query — `explain` Method

```
explain(): Unit (1)
explain(extended: Boolean): Unit
```

1. Turns the `extended` flag off (`false`)

`explain` simply prints out `explainInternal` to the standard output.

Note	<code>explain</code> is used when...FIXME
------	---

## `explainInternal` Method

```
explainInternal(extended: Boolean): String
```

`explainInternal` ...FIXME

	<code>explainInternal</code> is used when:
Note	<ul style="list-style-type: none"> <li>• <code>StreamExecution</code> is requested to <a href="#">explain a streaming query</a></li> <li>• <code>StreamingQueryWrapper</code> is requested to <a href="#">explainInternal</a></li> </ul>

## Stopping Streaming Sources and Readers — `stopSources` Method

```
stopSources(): Unit
```

`stopSources` requests every [streaming source](#) (in the [streaming query](#)) to [stop](#).

In case of an non-fatal exception, `stopSources` prints out the following WARN message to the logs:

```
Failed to stop streaming source: [source]. Resources may have leaked.
```

	<code>stopSources</code> is used when:
Note	<ul style="list-style-type: none"> <li>• <code>StreamExecution</code> is requested to <a href="#">run stream processing</a> (and <a href="#">terminates successfully or not</a>)</li> <li>• <code>ContinuousExecution</code> is requested to <a href="#">run the streaming query in continuous mode</a> (and <a href="#">terminates</a>)</li> </ul>

## Running Stream Processing — `runStream` Internal Method

```
runStream(): Unit
```

`runStream` simply prepares the environment to [execute the activated streaming query](#).

Note

`runStream` is used exclusively when the `stream execution thread` is requested to [start](#) (when `DataStreamWriter` is requested to [start an execution of the streaming query](#)).

Internally, `runStream` sets the job group (to all the Spark jobs started by this thread) as follows:

- `runId` for the job group ID
- `getBatchDescriptionString` for the job group description (to display in web UI)
- `interruptOnCancel` flag on

Note

`runStream` uses the `SparkSession` to access `SparkContext` and assign the job group id.

Read up on [SparkContext.setJobGroup](#) method in [The Internals of Apache Spark](#) book.

`runStream` sets `sql.streaming.queryId` local property to `id`.

`runStream` requests the `MetricsSystem` to register the `MetricsReporter` when `spark.sql.streaming.metricsEnabled` configuration property is on (default: off / `false` ).

`runStream` notifies `StreamingQueryListeners` that the streaming query has been started (by [posting](#) a new `QueryStartedEvent` event with `id`, `runId`, and `name`).

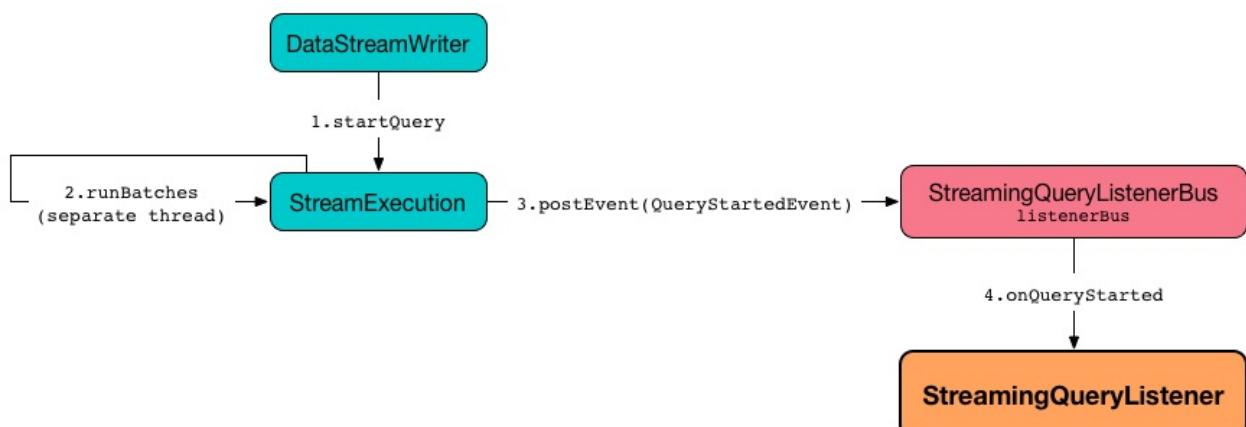


Figure 5. `StreamingQueryListener` Notified about Query's Start (`onQueryStarted`)

`runStream` unblocks the [main starting thread](#) (by decrementing the count of the `startLatch` that when `0` lets the starting thread continue).

Caution	FIXME A picture with two parallel lanes for the starting thread and daemon one for the query.
---------	---

`runStream` [updates the status message](#) to be **Initializing sources**.

`runStream` initializes the [analyzed logical plan](#).

Note	The <a href="#">analyzed logical plan</a> is a lazy value in Scala and is initialized when requested the very first time.
------	---

`runStream` disables **adaptive query execution** and **cost-based join optimization** (by turning `spark.sql.adaptive.enabled` and `spark.sql.cbo.enabled` configuration properties off, respectively).

`runStream` creates a new "zero" [OffsetSeqMetadata](#).

(Only when in **INITIALIZING** state) `runStream` enters **ACTIVE** state:

- Decrement the count of `initializationLatch`
- Executes the activated streaming query (which is different per [StreamExecution](#), i.e. [ContinuousExecution](#) or [MicroBatchExecution](#)).

Note	<code>runBatches</code> does the main work only when first started (i.e. when <code>state</code> is <b>INITIALIZING</b> ).
------	--

`runStream` ...FIXME (describe the failed and stop states)

Once [TriggerExecutor](#) has finished executing batches, `runBatches` [updates the status message](#) to **Stopped**.

Note	<a href="#">TriggerExecutor</a> finishes executing batches when <code>batch runner</code> returns whether the streaming query is stopped or not (which is when the internal <code>state</code> is not <b>TERMINATED</b> ).
------	--

Caution	FIXME Describe <code>catch</code> block for exception handling
---------	--

## Running Stream Processing — `finally` Block

`runStream` releases the `startLatch` and `initializationLatch` locks.

`runStream` [stopSources](#).

`runStream` sets the `state` to **TERMINATED**.

`runStream` sets the `StreamingQueryStatus` with the `isTriggerActive` and `isDataAvailable` flags off ( `false` ).

`runStream` removes the `stream metrics reporter` from the application's `MetricsSystem` .

`runStream` requests the `StreamingQueryManager` to handle termination of a streaming query.

`runStream` creates a new `QueryTerminatedEvent` (with the `id` and `run id` of the streaming query) and `posts it`.

With the `deleteCheckpointOnStop` flag enabled and no `StreamingQueryException` reported, `runStream` deletes the `checkpoint directory` recursively.

In the end, `runStream` releases the `terminationLatch` lock.

## TriggerExecutor's Batch Runner

**Batch Runner** (aka `batchRunner` ) is an executable block executed by `TriggerExecutor` in `runBatches`.

`batchRunner` starts trigger calculation.

As long as the query is not stopped (i.e. `state` is not `TERMINATED` ), `batchRunner` executes the streaming batch for the trigger.

In `triggerExecution` time-tracking section, `runBatches` branches off per `currentBatchId`.

Table 4. Current Batch Execution per currentBatchId

<code>currentBatchId &lt; 0</code>	<code>currentBatchId &gt;= 0</code>
<ol style="list-style-type: none"> <li>1. <code>populateStartOffsets</code></li> <li>2. Setting Job Description as <code>getBatchDescriptionString</code></li> </ol> <pre>DEBUG Stream running from [committedOffsets] to [availableOffsets]</pre>	<ol style="list-style-type: none"> <li>1. Constructing the next streaming micro-batch</li> </ol>

If there is `data available` in the sources, `batchRunner` marks `currentStatus` with `isDataAvailable` enabled.

**Note**

You can check out the status of a streaming query using `status` method.

```
scala> spark.streams.active(0).status
res1: org.apache.spark.sql.streaming.StreamingQueryStatus =
{
  "message" : "Waiting for next trigger",
  "isDataAvailable" : false,
  "isTriggerActive" : false
}
```

`batchRunner` then updates the status message to **Processing new data** and runs the current streaming batch.

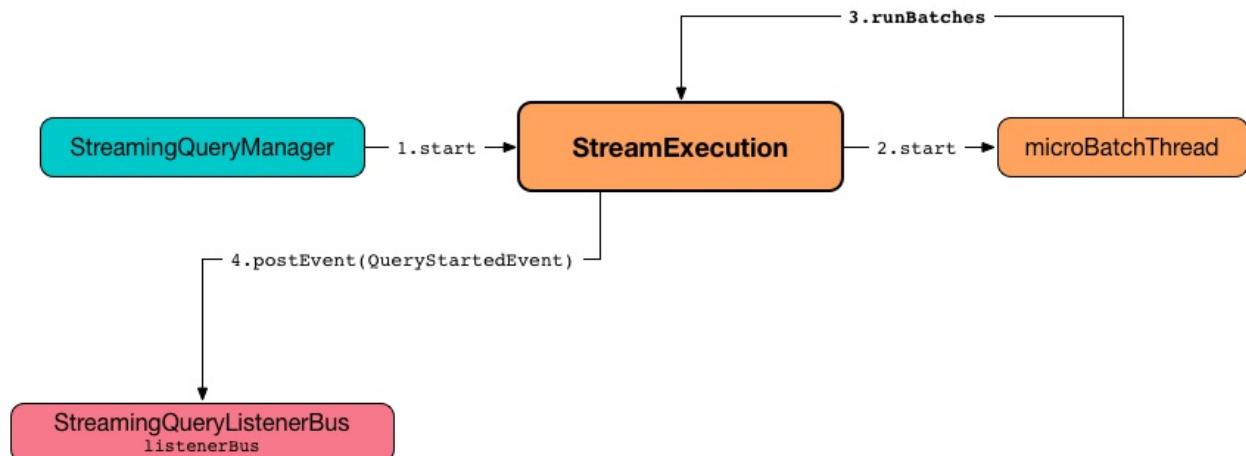


Figure 6. StreamExecution's Running Batches (on Execution Thread)

After **triggerExecution** section has finished, `batchRunner` **finishes the streaming batch for the trigger** (and collects query execution statistics).

When there was **data available** in the sources, `batchRunner` updates committed offsets (by adding the **current batch id** to **BatchCommitLog** and adding **availableOffsets** to **committedOffsets**).

You should see the following DEBUG message in the logs:

```
DEBUG batch $currentBatchId committed
```

`batchRunner` increments the **current batch id** and sets the job description for all the following Spark jobs to **include the new batch id**.

When no **data was available** in the sources to process, `batchRunner` does the following:

1. Marks `currentStatus` with `isDataAvailable` disabled
2. Updates the status message to **Waiting for data to arrive**
3. Sleeps the current thread for `pollingDelayMs` milliseconds.

`batchRunner` updates the status message to **Waiting for next trigger** and returns whether the query is currently active or not (so `TriggerExecutor` can decide whether to finish executing the batches or not)

## Starting Streaming Query (on Stream Execution Thread) — `start` Method

```
start(): Unit
```

When called, `start` prints out the following INFO message to the logs:

```
Starting [prettyIdString]. Use [resolvedCheckpointRoot] to store the query checkpoint.
```

`start` then starts the **stream execution thread** (as a daemon thread).

Note	<code>start</code> uses Java's <code>java.lang.Thread.start</code> to run the streaming query on a separate execution thread.
------	---

Note	When started, a streaming query runs in its own execution thread on JVM.
------	--

In the end, `start` pauses the main thread (using the `startLatch` until `StreamExecution` is requested to **run the streaming query** that in turn sends a `QueryStartedEvent` to all streaming listeners followed by decrementing the count of the `startLatch`).

Note	<code>start</code> is used exclusively when <code>StreamingQueryManager</code> is requested to <b>start a streaming query</b> (when <code>DataStreamWriter</code> is requested to <b>start an execution of the streaming query</b> ).
------	---

## Path to Checkpoint Directory — `checkpointFile` Internal Method

```
checkpointFile(name: String): String
```

`checkpointFile` gives the path of a directory with `name` in **checkpoint directory**.

Note	<code>checkpointFile</code> uses Hadoop's <code>org.apache.hadoop.fs.Path</code> .
------	--

**Note**

`checkpointFile` is used for `streamMetadata`, `OffsetSeqLog`, `BatchCommitLog`, and `lastExecution` (for `runBatch`).

## Posting StreamingQueryListener Event — `postEvent` Method

```
postEvent(event: StreamingQueryListener.Event): Unit
```

**Note**

`postEvent` is a part of [ProgressReporter Contract](#).

`postEvent` simply requests the `StreamingQueryManager` to `post` the input event (to the `StreamingQueryListenerBus` in the current `sparkSession` ).

**Note**

`postEvent` uses `SparkSession` to access the current `StreamingQueryManager`.

**Note**

`postEvent` is used when:

- `ProgressReporter` reports update progress (while finishing a trigger)
- `StreamExecution` runs streaming batches (and announces starting a streaming query by posting a `QueryStartedEvent` and query termination by posting a `QueryTerminatedEvent`)

## Waiting Until No New Data Available in Sources or Query Has Been Terminated — `processAllAvailable` Method

```
processAllAvailable(): Unit
```

**Note**

`processAllAvailable` is a part of [StreamingQuery Contract](#).

`processAllAvailable` reports the `StreamingQueryException` if reported (and returns immediately).

**Note**

`streamDeathCause` is reported exclusively when `streamExecution` is requested to run stream execution (that terminated with an exception).

`processAllAvailable` returns immediately when `streamExecution` is no longer `active` (in `TERMINATED` state).

`processAllAvailable` acquires a lock on the `awaitProgressLock` and turns the `noNewData` internal flag off (`false` ).

`processAllAvailable` keeps polling with 10-second pauses (locked on `awaitProgressLockCondition`) until `noNewData` flag is turned on (`true`) or `StreamExecution` is no longer `active` (in `TERMINATED` state).

Note	The 10-second pause is hardcoded and cannot be changed.
------	---

In the end, `processAllAvailable` releases `awaitProgressLock` lock.

`processAllAvailable` throws an `IllegalStateException` when executed on the `stream execution thread`:

```
Cannot wait for a query state from the same thread that is running the query
```

## Stream Execution Thread— `queryExecutionThread` Property

```
queryExecutionThread: QueryExecutionThread
```

`queryExecutionThread` is a Java thread of execution (`java.util.Thread`) that runs a streaming query.

`queryExecutionThread` is started (as a daemon thread) when `StreamExecution` is requested to `start`. At that time, `start` prints out the following INFO message to the logs (with the `prettyIdString` and the `resolvedCheckpointRoot`):

```
Starting [prettyIdString]. Use [resolvedCheckpointRoot] to store the query checkpoint.
```

When started, `queryExecutionThread` sets the call site and runs the streaming query.

`queryExecutionThread` uses the name **stream execution thread for [id]** (that uses `prettyIdString` for the id, i.e. `queryName [id = [id], runId = [runId]]` ).

`queryExecutionThread` is a `QueryExecutionThread` that is a custom `UninterruptibleThread` from Apache Spark with `runUninterruptibly` method for running a block of code without being interrupted by `Thread.interrupt()`.

	Use Java's <code>jconsole</code> or <code>jstack</code> to monitor stream execution threads.
--	--

Tip	\$ jstack <driver-pid>   grep -e "stream execution thread" "stream execution thread for kafka-topic1 [id =...]
-----	---

## Internal String Representation — `toDebugString` Internal Method

```
toDebugString(includeLogicalPlan: Boolean): String
```

`toDebugString` ...FIXME

Note	<code>toDebugString</code> is used exclusively when <code>streamExecution</code> is requested to run stream processing (and an exception is caught).
------	--

## Current Batch Metadata (Event-Time Watermark and Timestamp) — `offsetSeqMetadata` Internal Property

```
offsetSeqMetadata: OffsetSeqMetadata
```

`offsetSeqMetadata` is a [OffsetSeqMetadata](#).

Note	<code>offsetSeqMetadata</code> is part of the <a href="#">ProgressReporter Contract</a> to hold the current event-time watermark and timestamp.
------	---

`offsetSeqMetadata` is used to create an [IncrementalExecution](#) in the [queryPlanning](#) phase of the [MicroBatchExecution](#) and [ContinuousExecution](#) execution engines.

`offsetSeqMetadata` is initialized (with `0` for `batchWatermarkMs` and `batchTimestampMs`) when `StreamExecution` is requested to run stream processing.

`offsetSeqMetadata` is then updated (with the current event-time watermark and timestamp) when `MicroBatchExecution` is requested to construct the next streaming micro-batch.

Note	<code>MicroBatchExecution</code> uses the <a href="#">WatermarkTracker</a> for the current event-time watermark and the <a href="#">trigger clock</a> for the current batch timestamp.
------	--

`offsetSeqMetadata` is stored (*checkpointed*) in [walCommit phase](#) of `MicroBatchExecution` (and printed out as INFO message to the logs).

```
FIXME INFO message
```

`offsetSeqMetadata` is restored (*re-created*) from a checkpointed state when `MicroBatchExecution` is requested to populate start offsets.

## `isActive` Method

```
isActive: Boolean
```

**Note**

`isActive` is part of the [StreamingQuery Contract](#) to indicate whether a streaming query is active ( `true` ) or not ( `false` ).

`isActive` is enabled ( `true` ) as long as the [State](#) is not [TERMINATED](#).

**exception Method**

```
exception: Option[StreamingQueryException]
```

**Note**

`exception` is part of the [StreamingQuery Contract](#) to indicate whether a streaming query...FIXME

`exception` ...FIXME

**Human-Readable HTML Description of Spark Jobs (for web UI) — `getBatchDescriptionString` Method**

```
getBatchDescriptionString: String
```

`getBatchDescriptionString` is a human-readable description (in HTML format) that uses the optional `name` if defined, the `id`, the `runId` and `batchDescription` that can be `init` (for the [current batch ID negative](#)) or the [current batch ID itself](#).

`getBatchDescriptionString` is of the following format:

```
[name]<br/>id = [id]<br/>runId = [runId]<br/>batch =  
[batchDescription]
```

Job Id (Job Group)	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
5 (3480750c-4278-a24f-fef55af53c4c)	rate2console id = 8a5a3532-7f8d-4019-9b3a-9e6506087346 runId = 3480750c-4278-a24f-fef55af53c4c batch = 1	2019/09/25 12:12:00	67 ms	1/1	16/16

Figure 7. Monitoring Streaming Query using web UI (Spark Jobs)

Note	<p><code>getBatchDescriptionString</code> is used when:</p> <ul style="list-style-type: none"> <li>• <code>MicroBatchExecution</code> stream execution engine is requested to <a href="#">run an activated streaming query</a> (as the job description of any Spark jobs triggered as part of query execution)</li> <li>• <code>StreamExecution</code> is requested to <a href="#">run stream processing</a> (as the job group description of any Spark jobs triggered as part of query execution)</li> </ul>
------	---

## No New Data Available — `noNewData` Internal Flag

`noNewData: Boolean`

`noNewData` is a flag that indicates that a batch has completed with no new data left and [processAllAvailable](#) could stop waiting till all streaming data is processed.

Default: `false`

Turned on (`true`) when:

- `MicroBatchExecution` stream execution engine is requested to [construct or skip the next streaming micro-batch](#) (while [skipping the next micro-batch](#))
- `ContinuousExecution` stream execution engine is requested to [addOffset](#)

Turned off (`false`) when:

- `MicroBatchExecution` stream execution engine is requested to [construct or skip the next streaming micro-batch](#) (right after the [walCommit](#) phase)
- `StreamExecution` is requested to [processAllAvailable](#)

## Internal Properties

Name	Description
<code>awaitProgressLock</code>	Java's fair reentrant mutual exclusion <a href="#">java.util.concurrent.locks.ReentrantLock</a> (that favors granting access to the longest-waiting thread under contention)
<code>awaitProgressLockCondition</code>	Lock
<code>callSite</code>	
	Current batch ID

currentBatchId	<ul style="list-style-type: none"> <li>Starts at <code>-1</code> when <code>StreamExecution</code> is created</li> <li><code>0</code> when <code>StreamExecution</code> populates start offsets (and <code>OffsetSeqLog</code> is empty, i.e. no offset files in offsets directory in checkpoint)</li> <li>Incremented when <code>StreamExecution</code> runs streaming batches and finishes a trigger that had data available from sources (right after committing the batch).</li> </ul>		
initializationLatch			
newData	<p><code>newData: Map[BaseStreamingSource, LogicalPlan]</code></p> <p>Registry of the streaming sources (in the logical query plan) that have new data available in the current batch. The new data is a streaming DataFrame .</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <tr> <td style="padding: 5px;">Note</td> <td style="padding: 5px;"><code>newData</code> is part of the ProgressReporter Contract.</td> </tr> </table> <p>Set exclusively when <code>StreamExecution</code> is requested to requests unprocessed data from streaming sources (while running a single streaming batch).</p> <p>Used exclusively when <code>StreamExecution</code> is requested to transform the logical plan (of the streaming query) to include the Sources and the MicroBatchReaders with new data (while running a single streaming batch).</p>	Note	<code>newData</code> is part of the ProgressReporter Contract.
Note	<code>newData</code> is part of the ProgressReporter Contract.		
pollingDelayMs	<p>Time delay before polling new data again when no data was available</p> <p>Set to <code>spark.sql.streaming.pollingDelay</code> Spark property.</p> <p>Used when <code>StreamExecution</code> has started running streaming batches (and no data was available to process in a trigger).</p>		
prettyIdString	<p>Pretty-identified string for identification in logs (with name if defined).</p> <pre>queryName [id = xyz, runId = abc] [id = xyz, runId = abc]</pre>		
startLatch	Java's <code>java.util.concurrent.CountDownLatch</code> with count 1 .		

	Used when <code>StreamExecution</code> is requested to <a href="#">start</a> to pause the main thread until <code>StreamExecution</code> was requested to <a href="#">run the streaming query</a> .
<code>streamDeathCause</code>	<code>StreamingQueryException</code>
<code>streamMetrics</code>	<p><code>MetricsReporter</code> with <code>spark.streaming.[name or id]</code> source name</p> <p>Uses <code>name</code> if defined (can be <code>null</code>) or falls back to <code>id</code></p>
<code>uniqueSources</code>	<p>Unique <a href="#">streaming sources</a> (after being collected as <code>StreamingExecutionRelation</code> from the <a href="#">logical query plan</a>).</p> <div style="border: 1px solid black; padding: 10px;"> <p><b>Note</b> <code>StreamingExecutionRelation</code> is a leaf logical operator (i.e. <code>LogicalPlan</code>) that represents a streaming data source (and corresponds to a single <code>StreamingRelation</code> in <a href="#">analyzed logical query plan</a> of a streaming Dataset).</p> </div> <p>Used when <code>StreamExecution</code> :</p> <ul style="list-style-type: none"> <li>• <a href="#">Constructs the next streaming micro-batch</a> (and gets new offsets for every streaming data source)</li> <li>• <a href="#">Stops all streaming data sources</a></li> </ul>

# StreamingQueryWrapper — Serializable StreamExecution

StreamingQueryWrapper is a serializable interface of a StreamExecution.

Demo: Any Streaming Query is StreamingQueryWrapper

```
import org.apache.spark.sql.execution.streaming.StreamingQueryWrapper
val query = spark
  .readStream
  .format("rate")
  .load
  .writeStream
  .format("memory")
  .queryName("rate2memory")
  .start
assert(query.isInstanceOf[StreamingQueryWrapper])
```

StreamingQueryWrapper has the same StreamExecution API and simply passes all the method calls along to the underlying StreamExecution.

StreamingQueryWrapper is created when StreamingQueryManager is requested to create a streaming query (when DataStreamWriter is requested to start an execution of the streaming query).

# TriggerExecutor

`TriggerExecutor` is the [interface](#) for **trigger executors** that `StreamExecution` uses to execute a batch runner.

Note

**Batch runner** is an executable code that is executed at regular intervals. It is also called a **trigger handler**.

```
package org.apache.spark.sql.execution.streaming

trait TriggerExecutor {
  def execute(batchRunner: () => Boolean): Unit
}
```

Note

`StreamExecution` reports a `IllegalStateException` when `TriggerExecutor` is different from the [two built-in implementations](#): `OneTimeExecutor` or `ProcessingTimeExecutor`.

Table 1. `TriggerExecutor`'s Available Implementations

TriggerExecutor	Description		
<code>OneTimeExecutor</code>	Executes <code>batchRunner</code> exactly once.		
<code>ProcessingTimeExecutor</code>	<p>Executes <code>batchRunner</code> at regular intervals (as defined using <code>ProcessingTime</code> and <code>DataStreamWriter.trigger</code> method).</p> <pre>ProcessingTimeExecutor(   processingTime: ProcessingTime,   clock: Clock = new SystemClock())</pre> <table border="1"> <tr> <td>Note</td> <td>Processing terminates when <code>batchRunner</code> returns <code>false</code>.</td> </tr> </table>	Note	Processing terminates when <code>batchRunner</code> returns <code>false</code> .
Note	Processing terminates when <code>batchRunner</code> returns <code>false</code> .		

## notifyBatchFallingBehind Method

Caution

FIXME

# IncrementalExecution — QueryExecution of Streaming Queries

`IncrementalExecution` is the `QueryExecution` of streaming queries.

Tip

Read up on [QueryExecution](#) in [The Internals of Spark SQL book](#).

`IncrementalExecution` is [created](#) (and becomes the `StreamExecution.lastExecution`) when:

- `MicroBatchExecution` is requested to [run a single streaming micro-batch](#) (in `queryPlanning` phase)
- `ContinuousExecution` is requested to [run a streaming query in continuous mode](#) (in `queryPlanning` phase)
- `Dataset.explain` operator is executed (on a streaming query)

`IncrementalExecution` uses the `statefulOperatorId` internal counter for the IDs of the stateful operators in the [optimized logical plan](#) (while applying the [preparations](#) rules) when requested to prepare the plan for execution (in `executedPlan` phase).

## Preparing Logical Plan (of Streaming Query) for Execution — `optimizedPlan` and `executedPlan` Phases of Query Execution

When requested for the optimized logical plan (of the [logical plan](#)), `IncrementalExecution` transforms `CurrentBatchTimestamp` and `ExpressionWithRandomSeed` expressions with the timestamp literal and new random seeds, respectively. When transforming `CurrentBatchTimestamp` expressions, `IncrementalExecution` prints out the following INFO message to the logs:

```
Current batch timestamp = [timestamp]
```

Once [created](#), `IncrementalExecution` is immediately executed (by the `MicroBatchExecution` and `ContinuousExecution` stream execution engines in the `queryPlanning` phase) and so the entire query execution pipeline is executed up to and including `executedPlan`. That means that the [extra planning strategies](#) and the [state preparation rule](#) have been applied at this point and the [streaming query](#) is ready for execution.

## Creating IncrementalExecution Instance

`IncrementalExecution` takes the following to be created:

- `SparkSession`
- `Logical plan ( LogicalPlan )`
- `OutputMode` (as specified using `DataStreamWriter.outputMode` method)
- `State checkpoint location`
- Run ID of a streaming query ( `UUID` )
- Batch ID
- `OffsetSeqMetadata`

## State Checkpoint Location (Directory)

When `created`, `IncrementalExecution` is given the `checkpoint location`.

For the two available execution engines ([MicroBatchExecution](#) and [ContinuousExecution](#)), the checkpoint location is actually `state` directory under the `checkpoint root directory`.

```

val queryName = "rate2memory"
val checkpointLocation = s"file:/tmp/checkpoint-$queryName"
val query = spark
  .readStream
  .format("rate")
  .load
  .writeStream
  .format("memory")
  .queryName(queryName)
  .option("checkpointLocation", checkpointLocation)
  .start

// Give the streaming query a moment (one micro-batch)
// So lastExecution is available for the checkpointLocation
import scala.concurrent.duration._
query.awaitTermination(1.second.toMillis)

import org.apache.spark.sql.execution.streaming.StreamingQueryWrapper
val stateCheckpointDir = query
  .asInstanceOf[StreamingQueryWrapper]
  .streamingQuery
  .lastExecution
  .checkpointLocation
val stateDir = s"$checkpointLocation/state"
assert(stateCheckpointDir equals stateDir)

```

State checkpoint location is used exclusively when `IncrementalExecution` is requested for the [state info of the next stateful operator](#) (when requested to optimize a streaming physical plan using the [state preparation rule](#) that creates the stateful physical operators:

[StateStoreSaveExec](#), [StateStoreRestoreExec](#), [StreamingDeduplicateExec](#), [FlatMapGroupsWithStateExec](#), [StreamingSymmetricHashJoinExec](#), and [StreamingGlobalLimitExec](#)).

## Number of State Stores (`spark.sql.shuffle.partitions`)

### — `numStateStores` Internal Property

`numStateStores: Int`

`numStateStores` is the **number of state stores** which corresponds to `spark.sql.shuffle.partitions` configuration property (default: `200` ).

Tip

Read up on [spark.sql.shuffle.partitions](#) configuration property (and the others) in [The Internals of Spark SQL book](#).

Internally, `numStateStores` requests the [OffsetSeqMetadata](#) for the `spark.sql.shuffle.partitions` configuration property (using the [streaming configuration](#)) or simply takes whatever was defined for the given [SparkSession](#) (default: `200` ).

`numStateStores` is initialized right when `IncrementalExecution` is [created](#).

`numStateStores` is used exclusively when `IncrementalExecution` is requested for the [state info of the next stateful operator](#) (when requested to optimize a streaming physical plan using the [state preparation rule](#) that creates the stateful physical operators: [StateStoreSaveExec](#), [StateStoreRestoreExec](#), [StreamingDeduplicateExec](#), [FlatMapGroupsWithStateExec](#), [StreamingSymmetricHashJoinExec](#), and [StreamingGlobalLimitExec](#)).

## Extra Planning Strategies for Streaming Queries

### — `planner` Property

`IncrementalExecution` uses a custom `SparkPlanner` with the following **extra planning strategies** to plan the [streaming query](#) for execution:

- [StreamingJoinStrategy](#)
- [StatefulAggregationStrategy](#)
- [FlatMapGroupsWithStateStrategy](#)

- [StreamingRelationStrategy](#)
- [StreamingDeduplicationStrategy](#)
- [StreamingGlobalLimitStrategy](#)

**Tip**Read up on [SparkPlanner](#) in [The Internals of Spark SQL](#) book.

## State Preparation Rule For Execution-Specific Configuration — `state` Property

```
state: Rule[SparkPlan]
```

`state` is a custom physical preparation rule ( `Rule[SparkPlan]` ) that can transform a streaming physical plan ( `SparkPlan` ) with the following physical operators:

- [StateStoreSaveExec](#) with any unary physical operator ( `UnaryExecNode` ) with a [StateStoreRestoreExec](#)
- [StreamingDuplicateExec](#)
- [FlatMapGroupsWithStateExec](#)
- [StreamingSymmetricHashJoinExec](#)
- [StreamingGlobalLimitExec](#)

`state` simply transforms the physical plan with the above physical operators and fills out the execution-specific configuration:

- [nextStatefulOperationStateInfo](#) for the state info
- [OutputMode](#)
- [batchWatermarkMs](#) (through the [OffsetSeqMetadata](#)) for the event-time watermark
- [batchTimestampMs](#) (through the [OffsetSeqMetadata](#)) for the current timestamp
- [getStateWatermarkPredicates](#) for the state watermark predicates (for [StreamingSymmetricHashJoinExec](#))

`state` rule is used (as part of the physical query optimizations) when [IncrementalExecution](#) is requested to [optimize \(prepare\) the physical plan of the streaming query](#) (once for [ContinuousExecution](#) and every trigger for [MicroBatchExecution](#) in their [queryPlanning](#) phases).

**Tip**Read up on [Physical Query Optimizations](#) in [The Internals of Spark SQL](#) book.

## nextStatefulOperationStateInfo Internal Method

```
nextStatefulOperationStateInfo(): StatefulOperatorStateInfo
```

`nextStatefulOperationStateInfo` simply creates a new `StatefulOperatorStateInfo` with the `state checkpoint location`, the `run ID` (of the streaming query), the `next statefulOperator ID`, the `current batch ID`, and the `number of state stores`.

Note

The only changing part of `StatefulOperatorStateInfo` across calls of the `nextStatefulOperationStateInfo` method is the the `next statefulOperator ID`.

All the other properties (the `state checkpoint location`, the `run ID`, the `current batch ID`, and the `number of state stores`) are the same within a single `IncrementalExecution` instance.

The only two properties that may ever change are the `run ID` (after a streaming query is restarted from the checkpoint) and the `current batch ID` (every micro-batch in `MicroBatchExecution` execution engine).

Note

`nextStatefulOperationStateInfo` is used exclusively when `IncrementalExecution` is requested to optimize a streaming physical plan using the `state preparation rule` (and creates the stateful physical operators: `StateStoreSaveExec`, `StateStoreRestoreExec`, `StreamingDeduplicateExec`, `FlatMapGroupsWithStateExec`, `StreamingSymmetricHashJoinExec`, and `StreamingGlobalLimitExec`).

## Checking Out Whether Last Execution Requires Another Non-Data Micro-Batch — `shouldRunAnotherBatch` Method

```
shouldRunAnotherBatch(newMetadata: OffsetSeqMetadata): Boolean
```

`shouldRunAnotherBatch` is positive (`true`) if there is at least one `StateStoreWriter` operator (in the `executedPlan physical query plan`) that requires another non-data batch (per the given `OffsetSeqMetadata` with the event-time watermark and the batch timestamp).

Otherwise, `shouldRunAnotherBatch` is negative (`false`).

Note

`shouldRunAnotherBatch` is used exclusively when `MicroBatchExecution` is requested to `construct the next streaming micro-batch` (and checks out whether the last batch execution requires another non-data batch).

## Demo: State Checkpoint Directory

```

// START: Only for easier debugging
// The state is then only for one partition
// which should make monitoring easier
import org.apache.spark.sql.internal.SQLConf.SHUFFLE_PARTITIONS
spark.sessionState.conf.setConf(SHUFFLE_PARTITIONS, 1)

assert(spark.sessionState.conf.numShufflePartitions == 1)
// END: Only for easier debugging

val counts = spark
  .readStream
  .format("rate")
  .load
  .groupBy(window($"timestamp", "5 seconds") as "group")
  .agg(count("value") as "value_count") // <-- creates an Aggregate logical operator
  .orderBy("group") // <-- makes for easier checking

assert(counts.isStreaming, "This should be a streaming query")

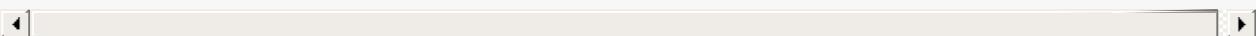
// Search for "checkpoint = <unknown>" in the following output
// Looks for StateStoreSave and StateStoreRestore
scala> counts.explain
== Physical Plan ==
*(5) Sort [group#5 ASC NULLS FIRST], true, 0
+- Exchange rangepartitioning(group#5 ASC NULLS FIRST, 1)
  +- *(4) HashAggregate(keys=[window#11], functions=[count(value#1L)])
    +- StateStoreSave [window#11], state info [ checkpoint = <unknown>, runId = 558b
f725-accb-487d-97eb-f790fa4a6138, opId = 0, ver = 0, numPartitions = 1], Append, 0, 2
    +- *(3) HashAggregate(keys=[window#11], functions=[merge_count(value#1L)])
      +- StateStoreRestore [window#11], state info [ checkpoint = <unknown>, run
Id = 558bf725-accb-487d-97eb-f790fa4a6138, opId = 0, ver = 0, numPartitions = 1], 2
      +- *(2) HashAggregate(keys=[window#11], functions=[merge_count(value#1L
)])
        +- Exchange hashpartitioning(window#11, 1)
          +- *(1) HashAggregate(keys=[window#11], functions=[partial_count(
value#1L)])
            +- *(1) Project [named_struct(start, precisetimestampconversion(((CASE WHEN (cast(CEIL((cast((precisetimestampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0)) as double) = (cast((precisetimestampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0)) THEN (CEIL((cast((precisetimestampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0)) + 1) ELSE CEIL((cast((precisetimestampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0)) END + 0) - 1) * 5000000) + 0), LongType, Tim
estampType), end, precisetimestampconversion(((CASE WHEN (cast(CEIL((cast((preciseti
mestampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0))
as double) = (cast((precisetimestampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0)) THEN (CEIL((cast((precisetimestampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0)) + 1) ELSE CEIL((cast((preciseti
mestampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0)) END + 0) - 1) * 5000000) + 5000000), LongType, TimestampType)) AS window#11, value#1L]
            +- *(1) Filter isnotnull(timestamp#0)
              +- StreamingRelation rate, [timestamp#0, value#1L]

```

```
// Start the query to access lastExecution that has the checkpoint resolved
import scala.concurrent.duration._
import org.apache.spark.sql.streaming.{OutputMode, Trigger}
val t = Trigger.ProcessingTime(1.hour) // should be enough time for exploration
val sq = counts
  .writeStream
  .format("console")
  .option("truncate", false)
  .option("checkpointLocation", "/tmp/spark-streams-state-checkpoint-root")
  .trigger(t)
  .outputMode(OutputMode.Complete)
  .start

// wait till the first batch which should happen right after start

import org.apache.spark.sql.execution.streaming._
val lastExecution = sq.asInstanceOf[StreamingQueryWrapper].streamingQuery.lastExecution
scala> println(lastExecution.checkpointLocation)
file:/tmp/spark-streams-state-checkpoint-root/state
```



# StreamingQueryListenerBus — Event Bus for Streaming Events

`StreamingQueryListenerBus` is an event bus (`ListenerBus[StreamingQueryListener, StreamingQueryListener.Event]`) for [dispatching streaming life-cycle events of active streaming queries](#) (that eventually are delivered to `StreamingQueryListeners`).

`StreamingQueryListenerBus` is created for `StreamingQueryManager` (once per `SparkSession`).

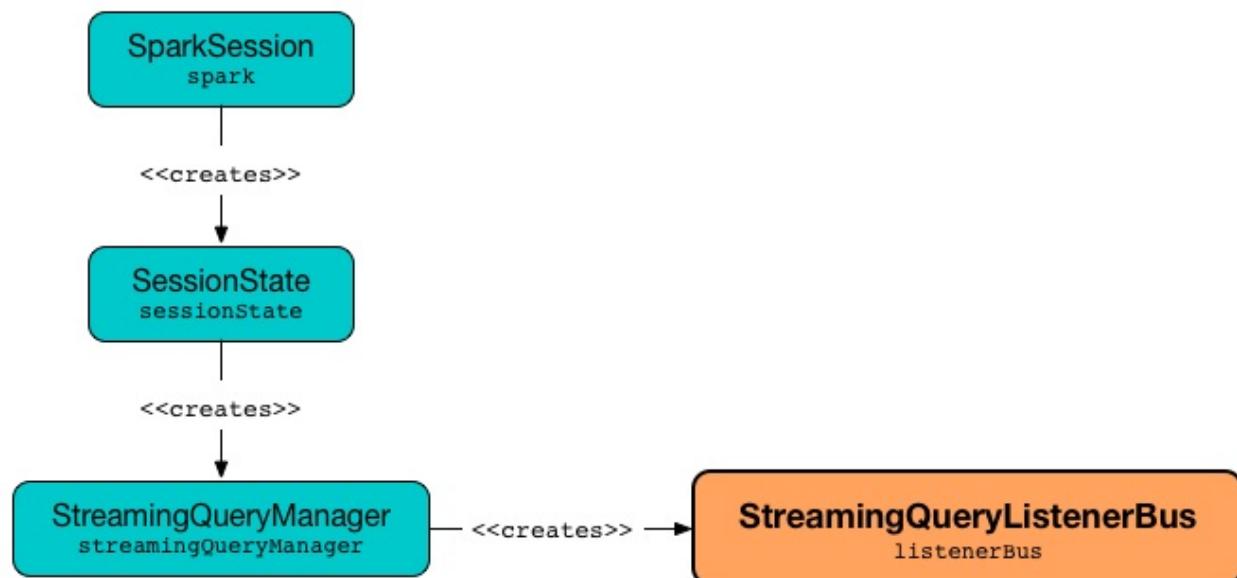


Figure 1. `StreamingQueryListenerBus` is Created Once In `SparkSession`

`StreamingQueryListenerBus` is also a `SparkListener` and registers itself with the `LiveListenerBus` (of the `SparkSession`) to [intercept QueryStartedEvents](#).

## Creating StreamingQueryListenerBus Instance

`StreamingQueryListenerBus` takes the following when created:

- `LiveListenerBus`

`StreamingQueryListenerBus` registers itself with the `LiveListenerBus`.

## Run IDs of Active Streaming Queries

```
activeQueryRunIds: HashSet[UUID]
```

`activeQueryRunIds` is an internal registry of [run IDs](#) of active streaming queries in the [SparkSession](#).

- A `runId` is added when `StreamingQueryListenerBus` is requested to [post a QueryStartedEvent](#)
- A `runId` is removed when `StreamingQueryListenerBus` is requested to [post a QueryTerminatedEvent](#)

`activeQueryRunIds` is used internally to [dispatch a streaming event](#) to a [StreamingQueryListener](#) (so the events gets sent out to streaming queries in the [SparkSession](#)).

## Posting Streaming Event to LiveListenerBus — `post` Method

```
post(event: StreamingQueryListener.Event): Unit
```

`post` simply posts the input `event` directly to the [LiveListenerBus](#) unless it is a [QueryStartedEvent](#).

For a [QueryStartedEvent](#), `post` adds the `runId` (of the streaming query that has been started) to the [activeQueryRunIds](#) internal registry first, posts the event to the [LiveListenerBus](#) and then [postToAll](#).

Note	<code>post</code> is used exclusively when <code>StreamingQueryManager</code> is requested to <a href="#">post a streaming event</a> .
------	--

## doPostEvent Method

```
doPostEvent(
  listener: StreamingQueryListener,
  event: StreamingQueryListener.Event): Unit
```

Note	<code>doPostEvent</code> is part of Spark Core's <code>ListenerBus</code> contract to post an event to the specified listener.
------	--

`doPostEvent` branches per the type of [StreamingQueryListener.Event](#):

- For a [QueryStartedEvent](#), requests the [StreamingQueryListener](#) to [onQueryStarted](#)
- For a [QueryProgressEvent](#), requests the [StreamingQueryListener](#) to [onQueryProgress](#)

- For a [QueryTerminatedEvent](#), requests the [StreamingQueryListener](#) to [onQueryTerminated](#)

For any other event, `doPostEvent` simply does nothing (*swallows it*).

## postToAll Method

```
postToAll(event: Event): Unit
```

Note

`postToAll` is part of Spark Core's `ListenerBus` contract to post an event to all registered listeners.

`postToAll` first requests the parent `ListenerBus` to post the event to all registered listeners.

For a [QueryTerminatedEvent](#), `postToAll` simply removes the `runId` (of the streaming query that has been terminated) from the [activeQueryRunIds](#) internal registry.

# StreamMetadata

`StreamMetadata` is a metadata associated with a [StreamingQuery](#) (indirectly through [StreamExecution](#)).

`StreamMetadata` takes an ID to be created.

`StreamMetadata` is [created](#) exclusively when [StreamExecution](#) is created (with a randomly-generated 128-bit universally unique identifier (UUID)).

`StreamMetadata` can be [persisted](#) to and [unpersisted](#) from a JSON file. `StreamMetadata` uses [json4s-jackson](#) library for JSON persistence.

```
import org.apache.spark.sql.execution.streaming.StreamMetadata
import org.apache.hadoop.fs.Path
val metadataPath = new Path("metadata")

scala> :type spark
org.apache.spark.sql.SparkSession

val hadoopConf = spark.sessionState.newHadoopConf()
val sm = StreamMetadata.read(metadataPath, hadoopConf)

scala> :type sm
Option[org.apache.spark.sql.execution.streaming.StreamMetadata]
```

## Unpersisting StreamMetadata (from JSON File) — `read` Object Method

```
read(
  metadataFile: Path,
  hadoopConf: Configuration): Option[StreamMetadata]
```

`read` unpersists `StreamMetadata` from the given `metadataFile` file if available.

`read` returns a `StreamMetadata` if the metadata file was available and the content could be read in JSON format. Otherwise, `read` returns `None`.

Note	<code>read</code> uses <code>org.json4s.jackson.Serialization.read</code> for JSON deserialization.
------	---

Note	<code>read</code> is used exclusively when <code>StreamExecution</code> is <a href="#">created</a> (and tries to read the <code>metadata</code> checkpoint file).
------	---

## Persisting Metadata — `write` Object Method

```
write(  
  metadata: StreamMetadata,  
  metadataFile: Path,  
  hadoopConf: Configuration): Unit
```

`write` persists the given `StreamMetadata` to the given `metadataFile` file in JSON format.

**Note** `write` uses `org.json4s.jackson.Serialization.write` for JSON serialization.

**Note** `write` is used exclusively when `streamExecution` is [created](#) (and the metadata checkpoint file is not available).

# EventTimeWatermark Unary Logical Operator — Streaming Watermark

`EventTimeWatermark` is a unary logical operator that is created to represent `Dataset.withWatermark` operator in a logical query plan of a streaming query.

Note	<p>A unary logical operator (<code>UnaryNode</code>) is a logical operator with a single <code>child</code> logical operator.</p> <p>Read up on <a href="#">UnaryNode</a> (and logical operators in general) in <a href="#">The Internals of Spark SQL book</a>.</p>
------	--

When requested for the `output attributes`, `EventTimeWatermark` logical operator goes over the output attributes of the `child` logical operator to find the matching attribute based on the `eventTime` attribute and updates it to include `spark.watermarkDelayMs` metadata key with the `watermark delay` interval (converted to milliseconds).

`EventTimeWatermark` is resolved (*planned*) to `EventTimeWatermarkExec` physical operator in `StatefulAggregationStrategy` execution planning strategy.

Note	<p><code>EliminateEventTimeWatermark</code> logical optimization rule (i.e. <code>Rule[LogicalPlan]</code> ) removes <code>EventTimeWatermark</code> logical operator from a logical plan if the <code>child</code> logical operator is streaming, i.e. when <code>Dataset.withWatermark</code> operator is used on a batch query.</p> <pre> val logs = spark.     read. // &lt;-- batch non-streaming query that makes `EliminateEventTimeWatermark` applicable     format("text").     load("logs")  // logs is a batch Dataset assert(!logs.isStreaming)  val q = logs.     withWatermark(eventTime = "timestamp", delayThreshold = "30 seconds") // &lt;-- EventTimeWatermark scala&gt; println(q.queryExecution.logical.numberedTreeString) // &lt;-- no EventTimeWatermark as it was removed immediately 00 Relation[value#0] text </pre>
------	---

## Creating EventTimeWatermark Instance

`EventTimeWatermark` takes the following to be created:

- Watermark column (`Attribute`)
- Watermark delay (`CalendarInterval`)

- Child logical operator ( `LogicalPlan` )

## Output Schema — `output` Property

```
output: Seq[Attribute]
```

Note	<code>output</code> is part of the <code>QueryPlan</code> Contract to describe the attributes of (the schema of) the output.
------	--

`output` finds `eventTime` column in the output schema of the `child` logical operator and updates the `Metadata` of the column with `spark.watermarkDelayMs` key and the milliseconds for the delay.

`output` removes `spark.watermarkDelayMs` key from the other columns.

```
// FIXME How to access/show the eventTime column with the metadata updated to include
spark.watermarkDelayMs?
import org.apache.spark.sql.catalyst.plans.logical.EventTimeWatermark
val etw = q.queryExecution.logical.asInstanceOf[EventTimeWatermark]
scala> etw.output.toStructType.printTreeString
root
|-- timestamp: timestamp (nullable = true)
|-- value: long (nullable = true)
```

## Watermark Metadata (Marker) — `spark.watermarkDelayMs` Metadata Key

`spark.watermarkDelayMs` metadata key is used to mark one of the `output attributes` as the **watermark attribute** (`eventTime watermark`).

## Converting Human-Friendly CalendarInterval to Milliseconds — `getDelayMs` Object Method

```
getDelayMs(
  delay: CalendarInterval): Long
```

`getDelayMs` ...FIXME

Note	<code>getDelayMs</code> is used when...FIXME
------	--



# FlatMapGroupsWithState Unary Logical Operator

`FlatMapGroupsWithState` is a unary logical operator that is [created](#) to represent the following operators in a logical query plan of a streaming query:

- [KeyValueGroupedDataset.mapGroupsWithState](#)
- [KeyValueGroupedDataset.flatMapGroupsWithState](#)

Note	A unary logical operator ( <code>UnaryNode</code> ) is a logical operator with a single <a href="#">child</a> logical operator.  Read up on <a href="#">UnaryNode</a> (and logical operators in general) in <a href="#">The Internals of Spark SQL book</a> .
------	---

`FlatMapGroupsWithState` is resolved (*planned*) to:

- [FlatMapGroupsWithStateExec](#) unary physical operator for streaming datasets (in [FlatMapGroupsWithStateStrategy](#) execution planning strategy)
- [MapGroupsExec](#) physical operator for batch datasets (in [BasicOperators](#) execution planning strategy)

## Creating `SerializeFromObject` with `FlatMapGroupsWithState` — `apply` Factory Method

```
apply[K: Encoder, V: Encoder, S: Encoder, U: Encoder](
  func: (Any, Iterator[Any], LogicalGroupState[Any]) => Iterator[Any],
  groupingAttributes: Seq[Attribute],
  dataAttributes: Seq[Attribute],
  outputMode: OutputMode,
  isMapGroupsWithState: Boolean,
  timeout: GroupStateTimeout,
  child: LogicalPlan): LogicalPlan
```

`apply` [creates](#) a `SerializeFromObject` logical operator with a `FlatMapGroupsWithState` as its child logical operator.

Internally, `apply` [creates](#) `SerializeFromObject` object consumer (aka unary logical operator) with `FlatMapGroupsWithState` logical plan.

Internally, `apply` finds `ExpressionEncoder` for the type `s` and creates a `FlatMapGroupsWithState` with `UnresolvedDeserializer` for the types `k` and `v`.

In the end, `apply` creates a `SerializeFromObject` object consumer with the `FlatMapGroupsWithState`.

**Note** `apply` is used in [KeyValueGroupedDataset.flatMapGroupsWithState](#) operator.

## Creating FlatMapGroupsWithState Instance

`FlatMapGroupsWithState` takes the following to be created:

- State function (`(Any, Iterator[Any], LogicalGroupState[Any]) ⇒ Iterator[Any]`)
- Key deserializer Catalyst expression
- Value deserializer Catalyst expression
- Grouping attributes
- Data attributes
- Output object attribute
- State `ExpressionEncoder`
- [Output mode](#)
- `isMapGroupsWithState` flag (default: `false`)
- [GroupStateTimeout](#)
- Child logical operator (`LogicalPlan`)

# Deduplicate Unary Logical Operator

`Deduplicate` is a unary logical operator (i.e. `LogicalPlan`) that is [created](#) to represent `dropDuplicates` operator (that drops duplicate records for a given subset of columns).

`Deduplicate` has [streaming](#) flag enabled for streaming Datasets.

```
val uniqueRates = spark.  
  readStream.  
  format("rate").  
  load.  
  dropDuplicates("value") // <-- creates Deduplicate logical operator  
// Note the streaming flag  
scala> println(uniqueRates.queryExecution.logical.numberedTreeString)  
00 Deduplicate [value#33L], true // <-- streaming flag enabled  
01 +- StreamingRelation DataSource(org.apache.spark.sql.SparkSession@4785f176, rate, List  
(,), None, List(), None, Map(), None), rate, [timestamp#32, value#33L]
```

## Caution

FIXME Example with duplicates across batches to show that `Deduplicate` keeps state and [withWatermark](#) operator should also be used to limit how much is stored (to not cause OOM)

## Note

`UnsupportedOperationChecker` [ensures](#) that `dropDuplicates` operator is not used after the following code is not supported in Structured Streaming and results in an `AnalysisException`:

```
val counts = spark.  
  readStream.  
  format("rate").  
  load.  
  groupBy(window($"timestamp", "5 seconds") as "group").  
  agg(count("value") as "value_count").  
  dropDuplicates // <-- after groupBy  
  
import scala.concurrent.duration._  
import org.apache.spark.sql.streaming.{OutputMode, Trigger}  
val sq = counts.  
  writeStream.  
  format("console").  
  trigger(Trigger.ProcessingTime(10.seconds)).  
  outputMode(OutputMode.Complete).  
  start  
org.apache.spark.sql.AnalysisException: dropDuplicates is not supported after a
```

Note	<p><code>Deduplicate</code> logical operator is translated (aka <i>planned</i>) to:</p> <ul style="list-style-type: none"><li>• <code>StreamingDeduplicateExec</code> physical operator in <code>StreamingDeduplicationStrategy</code> execution planning strategy for streaming Datasets (aka <i>streaming plans</i>)</li><li>• <code>Aggregate</code> physical operator in <code>ReplaceDeduplicatewithAggregate</code> execution planning strategy for non-streaming/batch Datasets (aka <i>batch plans</i>)</li></ul>
------	---

The output schema of `Deduplicate` is exactly the `child`'s output schema.

## Creating Deduplicate Instance

`Deduplicate` takes the following when created:

- Attributes for keys
- Child logical operator (i.e. `LogicalPlan` )
- Flag whether the logical operator is for streaming (enabled) or batch (disabled) mode

## MemoryPlan Logical Operator

`MemoryPlan` is a leaf logical operator (i.e. `LogicalPlan`) that is used to query the data that has been written into a `MemorySink`. `MemoryPlan` is created when [starting continuous writing](#) (to a `MemorySink`).

Tip

See the example in [MemoryStream](#).

```
scala> intsOut.explain(true)
== Parsed Logical Plan ==
SubqueryAlias memstream
+- MemoryPlan org.apache.spark.sql.execution.streaming.MemorySink@481bf251, [value#21]

== Analyzed Logical Plan ==
value: int
SubqueryAlias memstream
+- MemoryPlan org.apache.spark.sql.execution.streaming.MemorySink@481bf251, [value#21]

== Optimized Logical Plan ==
MemoryPlan org.apache.spark.sql.execution.streaming.MemorySink@481bf251, [value#21]

== Physical Plan ==
LocalTableScan [value#21]
```

When executed, `MemoryPlan` is translated to `LocalTableScanExec` physical operator (similar to `LocalRelation` logical operator) in `BasicOperators` execution planning strategy.

# StreamingRelation Leaf Logical Operator for Streaming Source

`StreamingRelation` is a leaf logical operator (i.e. `LogicalPlan`) that represents a [streaming source](#) in a logical plan.

`StreamingRelation` is [created](#) when `DataStreamReader` is requested to load data from a [streaming source](#) and creates a streaming `Dataset`.



Figure 1. `StreamingRelation` Represents Streaming Source

```

val rate = spark.
  readStream.      // <-- creates a DataStreamReader
  format("rate").
  load("hello")   // <-- creates a StreamingRelation
scala> println(rate.queryExecution.logical.numberedTreeString)
00 StreamingRelation DataSource(org.apache.spark.sql.SparkSession@4e5dcc50,rate,List(),
  None,None,Map(path -> hello),None), rate, [timestamp#0, value#1L]
  
```



`isStreaming` flag is always enabled (i.e. `true`).

```

import org.apache.spark.sql.execution.streaming.StreamingRelation
val relation = rate.queryExecution.logical.asInstanceOf[StreamingRelation]
scala> relation.isStreaming
res1: Boolean = true
  
```

`toString` gives the [source name](#).

```

scala> println(relation)
rate
  
```

Note

`StreamingRelation` is [resolved](#) (aka *planned*) to [StreamingExecutionRelation](#) (right after `StreamExecution` starts running batches).

## Creating StreamingRelation for DataSource — `apply` Object Method

```

apply(dataSource: DataSource): StreamingRelation
  
```

`apply` creates a `StreamingRelation` for the given `DataSource` (that represents a streaming source).

**Note**

`apply` is used exclusively when `DataStreamReader` is requested for a `streaming DataFrame`.

## Creating StreamingRelation Instance

`StreamingRelation` takes the following when created:

- `DataSource`
- Short name of the streaming source
- Output attributes of the schema of the streaming source

# StreamingRelationV2 Leaf Logical Operator

`StreamingRelationV2` is a `MultiInstanceRelation` leaf logical operator that represents `MicroBatchReadSupport` or `ContinuousReadSupport` streaming data sources in a logical plan of a streaming query.

Tip

Read up on [Leaf logical operators](#) in [The Internals of Spark SQL](#) book.

`StreamingRelationV2` is [created](#) when:

- `DataStreamReader` is requested to "load" data as a `streaming DataFrame` for `MicroBatchReadSupport` and `ContinuousReadSupport` streaming data sources
- `ContinuousMemoryStream` is created

`isStreaming` flag is always enabled (i.e. `true`).

```
scala> :type sq
org.apache.spark.sql.DataFrame

import org.apache.spark.sql.execution.streaming.StreamingRelationV2
val relation = sq.queryExecution.logical.asInstanceOf[StreamingRelationV2]
assert(relation.isStreaming)
```

`StreamingRelationV2` is resolved (*replaced*) to the following leaf logical operators:

- `ContinuousExecutionRelation` when `ContinuousExecution` stream execution engine is requested for the [analyzed logical plan](#)
- `StreamingExecutionRelation` when `MicroBatchExecution` stream execution engine is requested for the [analyzed logical plan](#)

## Creating StreamingRelationV2 Instance

`StreamingRelationV2` takes the following to be created:

- `DataSourceV2`
- Name of the data source
- Options ( `Map[String, String]` )
- Output attributes ( `seq[Attribute]` )
- Optional `StreamingRelation`

- `SparkSession`

# StreamingExecutionRelation Leaf Logical Operator for Streaming Source At Execution

`StreamingExecutionRelation` is a leaf logical operator (i.e. `LogicalPlan`) that represents a [streaming source](#) in the logical query plan of a streaming `Dataset`.

The main use of `StreamingExecutionRelation` logical operator is to be a "placeholder" in a logical query plan that will be replaced with the real relation (with new data that has arrived since the last batch) or an empty `LocalRelation` when `StreamExecution` is requested to [transforming logical plan to include the Sources and MicroBatchReaders with new data](#).

`StreamingExecutionRelation` is [created](#) for a `StreamingRelation` in [analyzed logical query plan](#) (that is the execution representation of a streaming `Dataset`).

## Note

Right after `StreamExecution` [has started running streaming batches](#) it initializes the streaming sources by transforming the analyzed logical plan of the streaming `Dataset` so that every `StreamingRelation` logical operator is replaced by the corresponding `StreamingExecutionRelation`.

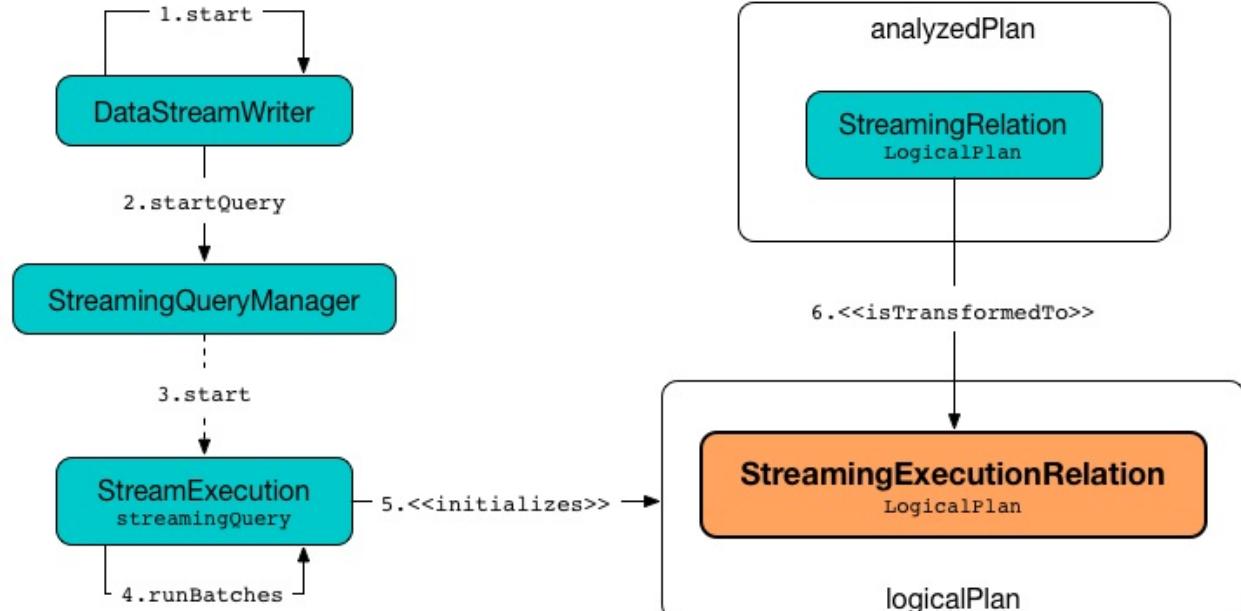


Figure 1. `StreamingExecutionRelation` Represents Streaming Source At Execution

## Note

`StreamingExecutionRelation` is also resolved (aka *planned*) to a `StreamingRelationExec` physical operator in `StreamingRelationStrategy` execution planning strategy only when [explaining](#) a streaming `Dataset`.

## Creating `StreamingExecutionRelation` Instance

`StreamingExecutionRelation` takes the following when created:

- Streaming source
- Output attributes

## Creating StreamingExecutionRelation (based on a Source)

### — apply Object Method

```
apply(source: Source): StreamingExecutionRelation
```

`apply` creates a `StreamingExecutionRelation` for the input `source` and with the attributes of the `schema` of the `source`.

Note	<code>apply</code> <i>seems</i> to be used for tests only.
------	--

# EventTimeWatermarkExec Unary Physical Operator

`EventTimeWatermarkExec` is a unary physical operator that represents [EventTimeWatermark](#) logical operator at execution time.

Note

A unary physical operator (`UnaryExecNode`) is a physical operator with a single [child](#) physical operator.

Read up on [UnaryExecNode](#) (and physical operators in general) in [The Internals of Spark SQL](#) book.

The purpose of the `EventTimeWatermarkExec` operator is to simply extract (*project*) the values of the [event-time watermark column](#) and add them directly to the [EventTimeStatsAccum](#) internal accumulator.

Note

Since the execution (data processing) happens on Spark executors, the only way to establish communication between the tasks (on the executors) and the driver is to use an accumulator.

Read up on [Accumulators](#) in [The Internals of Apache Spark](#) book.

`EventTimeWatermarkExec` uses [EventTimeStatsAccum](#) internal accumulator as a way to send the statistics (the maximum, minimum, average and update count) of the values in the [event-time watermark column](#) that is later used in:

- `ProgressReporter` for [creating execution statistics](#) for the most recent query execution (for monitoring the `max`, `min`, `avg`, and `watermark` event-time watermark statistics)
- `StreamExecution` to observe and possibly update event-time watermark when [constructing the next streaming batch](#).

`EventTimeWatermarkExec` is [created](#) exclusively when [StatefulAggregationStrategy](#) execution planning strategy is requested to plan a logical plan with [EventTimeWatermark](#) logical operators for execution.

Tip

Check out [Demo: Streaming Watermark with Aggregation in Append Output Mode](#) to deep dive into the internals of [Streaming Watermark](#).

## Creating EventTimeWatermarkExec Instance

`EventTimeWatermarkExec` takes the following to be created:

- **Event time column** - the column with the (event) time for event-time watermark
- Delay interval ( `CalendarInterval` )
- Child physical operator ( `SparkPlan` )

While `being created`, `EventTimewatermarkExec` registers the `EventTimeStatsAccum` internal accumulator (with the current `SparkContext` ).

## Executing Physical Operator (Generating RDD[InternalRow]) — `doExecute` Method

```
doExecute(): RDD[InternalRow]
```

Note	<code>doExecute</code> is part of <code>SparkPlan</code> Contract to generate the runtime representation of an physical operator as a distributed computation over internal binary rows on Apache Spark (i.e. <code>RDD[InternalRow]</code> ).
------	--

Internally, `doExecute` executes the `child` physical operator and maps over the partitions (using `RDD.mapPartitions` ).

`doExecute` creates an unsafe projection (one per partition) for the `column with the event time` in the output schema of the `child` physical operator. The unsafe projection is to extract event times from the (stream of) internal rows of the child physical operator.

For every row ( `InternalRow` ) per partition, `doExecute` requests the `eventTimeStats` accumulator to `add the event time`.

Note	The event time value is in seconds (not millis as the value is divided by <code>1000</code> ).
------	--

## Output Attributes (Schema) — `output` Property

```
output: Seq[Attribute]
```

Note	<code>output</code> is part of the <code>QueryPlan</code> Contract to describe the attributes of (the schema of) the output.
------	--

`output` requests the `child` physical operator for the output attributes to find the `event time column` and any other column with metadata that contains `spark.watermarkDelayMs` key.

For the `event time column`, `output` updates the metadata to include the `delay` interval for the `spark.watermarkDelayMs` key.

For any other column (not the event time column) with the `spark.watermarkDelayMs` key, `output` simply removes the key from the metadata.

```
// FIXME: Would be nice to have a demo. Anyone?
```

## Internal Properties

Name	Description				
<code>delayMs</code>	<p><b>Delay interval</b> - the <code>delay</code> interval in milliseconds</p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>EventTimeWatermarkExec</code> is requested for the <code>output</code> attributes</li> <li>• <code>WatermarkTracker</code> is requested to update the event-time watermark</li> </ul>				
<code>eventTimeStats</code>	<p><code>EventTimeStatsAccum</code> accumulator to accumulate <code>eventTime</code> values from every row in a streaming batch (when <code>EventTimeWatermarkExec</code> is executed).</p> <table border="1"> <tr> <td>Note</td><td><code>EventTimeStatsAccum</code> is a Spark accumulator of <code>EventTimeStats</code> from <code>Longs</code> (i.e. <code>AccumulatorV2[Long, EventTimeStats]</code> ).</td></tr> <tr> <td>Note</td><td>Every Spark accumulator has to be registered before use, and <code>eventTimeStats</code> is registered when <code>EventTimeWatermarkExec</code> is created.</td></tr> </table>	Note	<code>EventTimeStatsAccum</code> is a Spark accumulator of <code>EventTimeStats</code> from <code>Longs</code> (i.e. <code>AccumulatorV2[Long, EventTimeStats]</code> ).	Note	Every Spark accumulator has to be registered before use, and <code>eventTimeStats</code> is registered when <code>EventTimeWatermarkExec</code> is created.
Note	<code>EventTimeStatsAccum</code> is a Spark accumulator of <code>EventTimeStats</code> from <code>Longs</code> (i.e. <code>AccumulatorV2[Long, EventTimeStats]</code> ).				
Note	Every Spark accumulator has to be registered before use, and <code>eventTimeStats</code> is registered when <code>EventTimeWatermarkExec</code> is created.				

# FlatMapGroupsWithStateExec Unary Physical Operator

`FlatMapGroupsWithStateExec` is a unary physical operator that represents `FlatMapGroupsWithState` logical operator at execution time.

Note

A unary physical operator (`UnaryExecNode`) is a physical operator with a single `child` physical operator.

Read up on [UnaryExecNode](#) (and physical operators in general) in [The Internals of Spark SQL](#) book.

Note

`FlatMapGroupsWithState` unary logical operator represents `KeyValueGroupedDataset.mapGroupsWithState` and `KeyValueGroupedDataset.flatMapGroupsWithState` operators in a logical query plan.

`FlatMapGroupsWithStateExec` is [created](#) exclusively when `FlatMapGroupsWithStateStrategy` execution planning strategy is requested to plan a `FlatMapGroupsWithState` logical operator for execution.

`FlatMapGroupsWithStateExec` is an `ObjectProducerExec` physical operator and so produces a [single output object](#).

Tip

Read up on [ObjectProducerExec—Physical Operators With Single Object Output](#) in [The Internals of Spark SQL](#) book.

Tip

Check out [Demo: Internals of FlatMapGroupsWithStateExec Physical Operator](#).

Note

`FlatMapGroupsWithStateExec` is given an `OutputMode` when created, but it does not seem to be used at all. Check out the question [What's the purpose of OutputMode in flatMapGroupsWithState? How/where is it used?](#) on StackOverflow.

Tip

Enable `ALL` logging level for `org.apache.spark.sql.execution.streaming.FlatMapGroupsWithStateExec` to see what happens inside.

Add the following line to `conf/log4j.properties`:

```
log4j.logger.org.apache.spark.sql.execution.streaming.FlatMapGroupsWithStateExec=
```

Refer to [Logging](#).

## Creating FlatMapGroupsWithStateExec Instance

FlatMapGroupsWithStateExec takes the following to be created:

- **User-defined state function** that is applied to every group (of type `(Any, Iterator[Any], LogicalGroupState[Any]) → Iterator[Any]`)
- Key deserializer expression
- Value deserializer expression
- Grouping attributes (as used for grouping in [KeyValueGroupedDataset](#) for `mapGroupsWithState` or `flatMapGroupsWithState` operators)
- Data attributes
- Output object attribute (that is the reference to the single object field this operator outputs)
- [StatefulOperatorStateInfo](#)
- State encoder (`ExpressionEncoder[Any]`)
- State format version
- [OutputMode](#)
- [GroupStateTimeout](#)
- [Batch Processing Time](#)
- [Event-time watermark](#)
- Child physical operator

FlatMapGroupsWithStateExec initializes the [internal properties](#).

## Performance Metrics (SQLMetrics)

FlatMapGroupsWithStateExec uses the performance metrics of [StateStoreWriter](#).

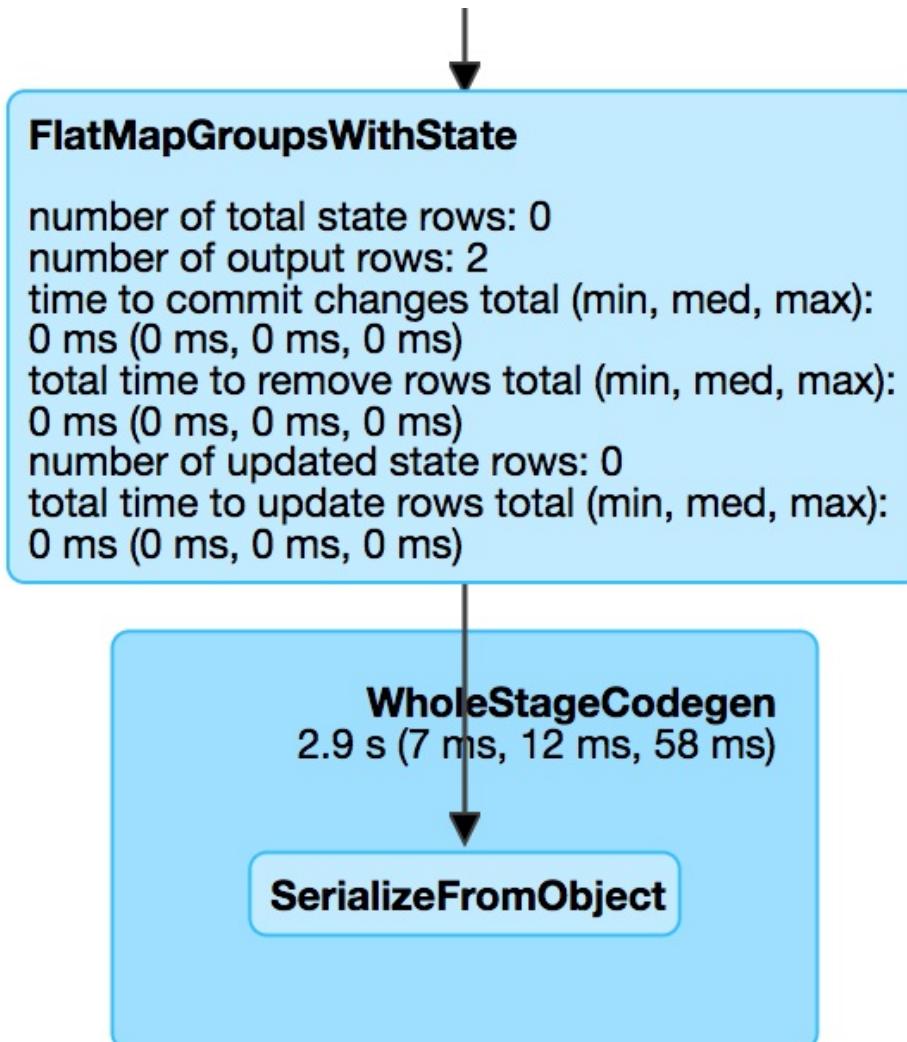


Figure 1. FlatMapGroupsWithStateExec in web UI (Details for Query)

## FlatMapGroupsWithStateExec as StateStoreWriter

`FlatMapGroupsWithStateExec` is a [stateful physical operator that can write to a state store](#)(and `MicroBatchExecution` requests [whether to run another batch or not](#) based on the `GroupStateTimeout`).

`FlatMapGroupsWithStateExec` uses the `GroupStateTimeout` (and possibly the updated `metadata`) when asked [whether to run another batch or not](#) (when `MicroBatchExecution` is requested to [construct the next streaming micro-batch](#) when requested to [run the activated streaming query](#)).

## FlatMapGroupsWithStateExec with Streaming Event-Time Watermark Support (WatermarkSupport)

`FlatMapGroupsWithStateExec` is a [physical operator that supports streaming event-time watermark](#).

`FlatMapGroupsWithStateExec` is given the [optional event time watermark](#) when created.

The [event-time watermark](#) is initially undefined (`None`) when planned to for execution (in [FlatMapGroupsWithStateStrategy](#) execution planning strategy).

**Note**

`FlatMapGroupsWithStateStrategy` converts [FlatMapGroupsWithState](#) unary logical operator to `FlatMapGroupsWithStateExec` physical operator with undefined [StatefulOperatorStateInfo](#), [batchTimestampMs](#), and [eventTimeWatermark](#).

The [event-time watermark](#) (with the [StatefulOperatorStateInfo](#) and the [batchTimestampMs](#)) is only defined to the [current event-time watermark](#) of the given [OffsetSeqMetadata](#) when [IncrementalExecution](#) query execution pipeline is requested to apply the [state](#) preparation rule (as part of the [preparations](#) rules).

**Note**

The [preparations](#) rules are executed (applied to a physical query plan) at the [executedPlan](#) phase of Structured Query Execution Pipeline to generate an optimized physical query plan ready for execution).

Read up on [Structured Query Execution Pipeline](#) in [The Internals of Spark SQL](#) book.

[IncrementalExecution](#) is used as the [lastExecution](#) of the available [streaming query execution engines](#). It is created in the [queryPlanning](#) phase (of the [MicroBatchExecution](#) and [ContinuousExecution](#) execution engines) based on the current [OffsetSeqMetadata](#).

**Note**

The [optional event-time watermark](#) can only be defined when the [state](#) preparation rule is executed which is at the [executedPlan](#) phase of Structured Query Execution Pipeline which is also part of the [queryPlanning](#) phase.

## FlatMapGroupsWithStateExec and StateManager — stateManager Property

`stateManager: StateManager`

While being created, `FlatMapGroupsWithStateExec` creates a [StateManager](#) (with the [state encoder](#) and the [isTimeoutEnabled](#) flag).

A `stateManager` is [created](#) per [state format version](#) that is given while creating a `FlatMapGroupsWithStateExec` (to choose between the [available implementations](#)).

The [state format version](#) is controlled by `spark.sql.streaming.flatMapGroupsWithState.stateFormatVersion` internal configuration property (default: `2` ).

Note	<code>StateManagerImplV2</code> is the default <code>StateManager</code> .
------	--

The `StateManager` is used exclusively when `FlatMapGroupsWithStateExec` physical operator is [executed](#) (to generate a recipe for a distributed computation as an `RDD[InternalRow]`) for the following:

- [State schema](#) (for the [value schema](#) of a `StateStoreRDD`)
- [State data for a key in a StateStore](#) while processing new data
- All state data (for all keys) in a `StateStore` while [processing timed-out state data](#)
- Removing the state for a key from a `StateStore` when [all rows have been processed](#)
- Persisting the state for a key in a `StateStore` when [all rows have been processed](#)

## keyExpressions Method

```
keyExpressions: Seq[Attribute]
```

Note	<code>keyExpressions</code> is part of the <a href="#">WatermarkSupport Contract</a> to...FIXME.
------	--

`keyExpressions` simply returns the [grouping attributes](#).

## Executing Physical Operator (Generating RDD[InternalRow]) — doExecute Method

```
doExecute(): RDD[InternalRow]
```

Note	<code>doExecute</code> is part of <code>SparkPlan</code> Contract to generate the runtime representation of an physical operator as a distributed computation over internal binary rows on Apache Spark (i.e. <code>RDD[InternalRow]</code> ).
------	--

`doExecute` first initializes the [metrics](#) (which happens on the driver).

`doExecute` then requests the [child](#) physical operator to execute and generate an `RDD[InternalRow]`.

`doExecute` uses [StateStoreOps](#) to create a `StateStoreRDD` with a `storeUpdateFunction` that does the following (for a partition):

1. Creates an [InputProcessor](#) for a given [StateStore](#)

2. (only when the `GroupStateTimeout` is `EventTimeTimeout`) Filters out late data based on the `event-time watermark`, i.e. rows from a given `Iterator[InternalRow]` that are older than the `event-time watermark` are excluded from the steps that follow
3. Requests the `InputProcessor` to create an iterator of a new data processed from the (possibly filtered) iterator
4. Requests the `InputProcessor` to create an iterator of a timed-out state data
5. Creates an iterator by concatenating the above iterators (with the new data processed first)
6. In the end, creates a `CompletionIterator` that executes a completion function (`completionFunction`) after it has successfully iterated through all the elements (i.e. when a client has consumed all the rows). The completion method requests the given `StateStore` to commit changes followed by setting the store-specific metrics.

## Checking Out Whether Last Batch Execution Requires Another Non-Data Batch or Not

### — `shouldRunAnotherBatch` Method

```
shouldRunAnotherBatch(newMetadata: OffsetSeqMetadata): Boolean
```

**Note**

`shouldRunAnotherBatch` is part of the `StateStoreWriter Contract` to indicate whether `MicroBatchExecution` should run another non-data batch (based on the updated `OffsetSeqMetadata` with the current event-time watermark and the batch timestamp).

`shouldRunAnotherBatch` uses the `GroupStateTimeout` as follows:

- With `EventTimeTimeout`, `shouldRunAnotherBatch` is positive (`true`) only when the `event-time watermark` is defined and is older (below) the `event-time watermark` of the given `OffsetSeqMetadata`
- With `NoTimeout` (and other `GroupStateTimeouts` if there were any), `shouldRunAnotherBatch` is always negative (`false`)
- With `ProcessingTimeTimeout`, `shouldRunAnotherBatch` is always positive (`true`)

## Internal Properties

Name	Description
<code>isTimeoutEnabled</code>	<p>Flag that says whether the <a href="#">GroupStateTimeout</a> is not <a href="#">NoTimeout</a></p> <p>Used when:</p> <ul style="list-style-type: none"> <li>• <code>FlatMapGroupsWithStateExec</code> is created (and creates the internal <a href="#">StateManager</a>)</li> <li>• <code>InputProcessor</code> is requested to <a href="#">processTimedOutState</a></li> </ul>
<code>stateAttributes</code>	
<code>stateDeserializer</code>	
<code>stateSerializer</code>	
<code>timestampTimeoutAttribute</code>	
<code>watermarkPresent</code>	<p>Flag that says whether the <a href="#">child</a> physical operator has a <a href="#">watermark attribute</a> (among the output attributes).</p> <p>Used exclusively when <code>InputProcessor</code> is requested to <a href="#">callFunctionAndUpdateState</a></p>

# StateStoreRestoreExec Unary Physical Operator — Restoring Streaming State From State Store

`StateStoreRestoreExec` is a unary physical operator that [restores](#) (reads) a streaming state from a [state store](#) (for the keys from the [child](#) physical operator).

Note	<p>A unary physical operator (<code>UnaryExecNode</code>) is a physical operator with a single <a href="#">child</a> physical operator.</p> <p>Read up on <a href="#">UnaryExecNode</a> (and physical operators in general) in <a href="#">The Internals of Spark SQL</a> book.</p>
------	---

`StateStoreRestoreExec` is [created](#) exclusively when [StatefulAggregationStrategy](#) execution planning strategy is requested to plan a [streaming aggregation](#) for execution ( [Aggregate](#) logical operators in the logical plan of a streaming query).

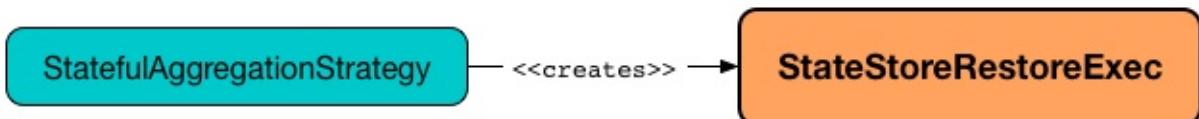


Figure 1. StateStoreRestoreExec and StatefulAggregationStrategy

The optional [StatefulOperatorStateInfo](#) is initially undefined (i.e. when `StateStoreRestoreExec` is [created](#)). `StateStoreRestoreExec` is updated to hold the streaming batch-specific execution property when `IncrementalExecution` [prepares a streaming physical plan for execution](#) (and [state](#) preparation rule is executed when `StreamExecution` [plans a streaming query](#) for a streaming batch).



Figure 2. StateStoreRestoreExec and IncrementalExecution

When [executed](#), `StateStoreRestoreExec` executes the [child](#) physical operator and [creates a StateStoreRDD to map over partitions](#) with `storeUpdateFunction` that restores the state for the keys in the input rows if available.

The output schema of `StateStoreRestoreExec` is exactly the [child](#)'s output schema.

The output partitioning of `StateStoreRestoreExec` is exactly the [child](#)'s output partitioning.

## Performance Metrics (SQLMetrics)

Key	Name (in UI)	Description
numOutputRows	number of output rows	The number of input rows from the <code>child</code> physical operator (for which <code>StateStoreRestoreExec</code> tried to find the state)

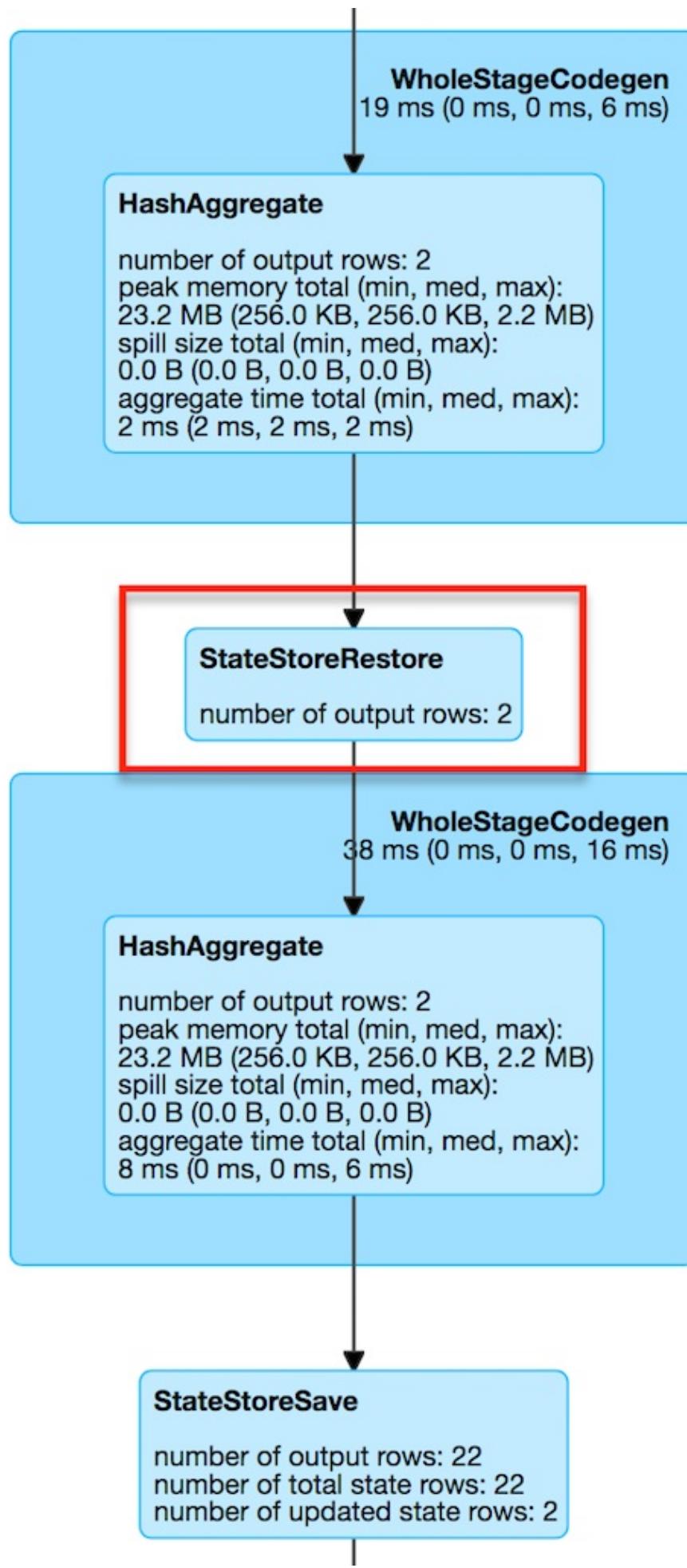


Figure 3. StateStoreRestoreExec in web UI (Details for Query)

## Creating StateStoreRestoreExec Instance

`StateStoreRestoreExec` takes the following to be created:

- **Key expressions**, i.e. Catalyst attributes for the grouping keys
- Optional `StatefulOperatorStateInfo` (default: `None`)
- Version of the state format (based on the `spark.sql.streaming.aggregation.stateFormatVersion` configuration property)
- Child physical operator (`SparkPlan`)

## StateStoreRestoreExec and StreamingAggregationStateManager — `stateManager` Property

```
stateManager: StreamingAggregationStateManager
```

`stateManager` is a `StreamingAggregationStateManager` that is created together with `StateStoreRestoreExec`.

The `StreamingAggregationStateManager` is created for the `keys`, the output schema of the child physical operator and the `version of the state format`.

The `StreamingAggregationStateManager` is used when `StateStoreRestoreExec` is requested to generate a recipe for a distributed computation (as a `RDD[InternalRow]`) for the following:

- Schema of the values in a state store
- Extracting the columns for the key from the input row
- Looking up the value of a key from a state store

## Executing Physical Operator (Generating RDD[InternalRow]) — `doExecute` Method

```
doExecute(): RDD[InternalRow]
```

## Note

`doExecute` is part of `SparkPlan` Contract to generate the runtime representation of an physical operator as a distributed computation over internal binary rows on Apache Spark (i.e. `RDD[InternalRow]` ).

Internally, `doExecute` executes `child` physical operator and creates a `StateStoreRDD` with `storeUpdateFunction` that does the following per `child` operator's RDD partition:

1. Generates an unsafe projection to access the key field (using `keyExpressions` and the output schema of `child` operator).
2. For every input row (as `InternalRow`)
  - Extracts the key from the row (using the unsafe projection above)
  - Gets the saved state in `StateStore` for the key if available (it might not be if the key appeared in the input the first time)
  - Increments `numOutputRows` metric (that in the end is the number of rows from the `child` operator)
  - Generates collection made up of the current row and possibly the state for the key if available

## Note

The number of rows from `stateStoreRestoreExec` is the number of rows from the `child` operator with additional rows for the saved state.

## Note

There is no way in `StateStoreRestoreExec` to find out how many rows had associated state available in a state store. You would have to use the corresponding `StateStoreSaveExec` operator's `metrics` (most likely `number of total state rows` but that could depend on the output mode).

# StateStoreSaveExec Unary Physical Operator — Saving Streaming State To State Store

`StateStoreSaveExec` is a unary physical operator that saves a streaming state to a state store with support for streaming watermark.

Note	A unary physical operator ( <code>UnaryExecNode</code> ) is a physical operator with a single child physical operator.  Read up on <a href="#">UnaryExecNode</a> (and physical operators in general) in <a href="#">The Internals of Spark SQL</a> book.
------	--

`StateStoreSaveExec` is created exclusively when `StatefulAggregationStrategy` execution planning strategy is requested to plan a streaming aggregation for execution (Aggregate logical operators in the logical plan of a streaming query).

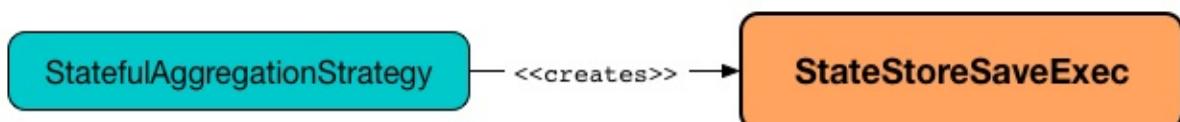


Figure 1. StateStoreSaveExec and StatefulAggregationStrategy

The optional properties, i.e. the `StatefulOperatorStateInfo`, the `output mode`, and the `event-time watermark`, are initially undefined when `StateStoreSaveExec` is created.

`StateStoreSaveExec` is updated to hold execution-specific configuration when `IncrementalExecution` is requested to prepare the logical plan (of a streaming query) for execution (when the state preparation rule is executed).



Figure 2. StateStoreSaveExec and IncrementalExecution

Note	Unlike <code>StateStoreRestoreExec</code> operator, <code>stateStoreSaveExec</code> takes output mode and event time watermark when created.
------	--

When executed, `StateStoreSaveExec` creates a `StateStoreRDD` to map over partitions with `storeUpdateFunction` that manages the `StateStore`.

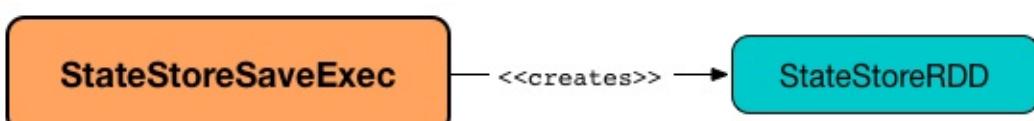


Figure 3. StateStoreSaveExec creates StateStoreRDD

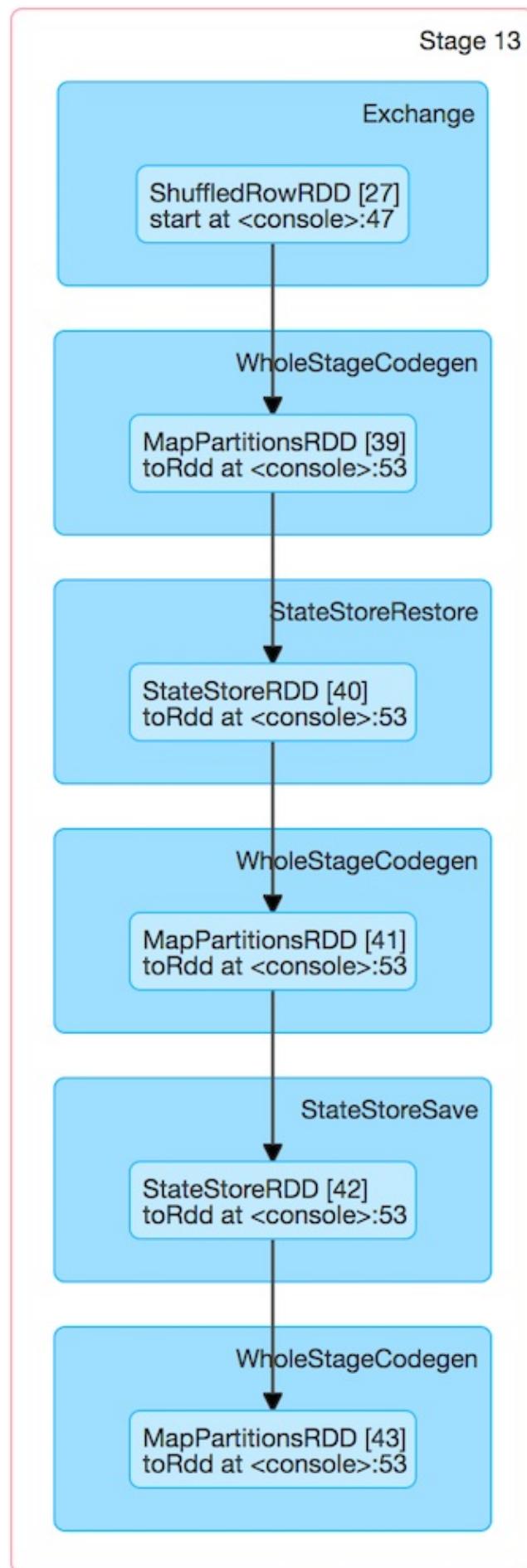


Figure 4. StateStoreSaveExec and StateStoreRDD (after streamingBatch.toRdd.count)

Note	The number of partitions of <b>StateStoreRDD</b> (and hence the number of Spark tasks) is what was defined for the <b>child</b> physical plan. There will be that many <code>stateStores</code> as there are partitions in <code>StateStoreRDD</code> .
------	--

Note	<code>StateStoreSaveExec</code> <b>behaves</b> differently per output mode.
------	---

When `executed`, `StateStoreSaveExec` executes the **child** physical operator and creates a **StateStoreRDD** (with `storeUpdateFunction` specific to the output mode).

The output schema of `StateStoreSaveExec` is exactly the **child**'s output schema.

The output partitioning of `StateStoreSaveExec` is exactly the **child**'s output partitioning.

`StateStoreRestoreExec` uses a **StreamingAggregationStateManager** (that is `created` for the `keyExpressions`, the output of the **child** physical operator and the `stateFormatVersion`).

Tip	<p>Enable <code>ALL</code> logging level for  <code>org.apache.spark.sql.execution.streaming.StateStoreSaveExec</code> to see what happens inside.</p> <p>Add the following line to <code>conf/log4j.properties</code> :</p> <pre>log4j.logger.org.apache.spark.sql.execution.streaming.StateStoreSaveExec=ALL</pre> <p>Refer to <a href="#">Logging</a>.</p>
-----	---

## Performance Metrics (SQLMetrics)

`StateStoreSaveExec` uses the performance metrics as other stateful physical operators that write to a state store.

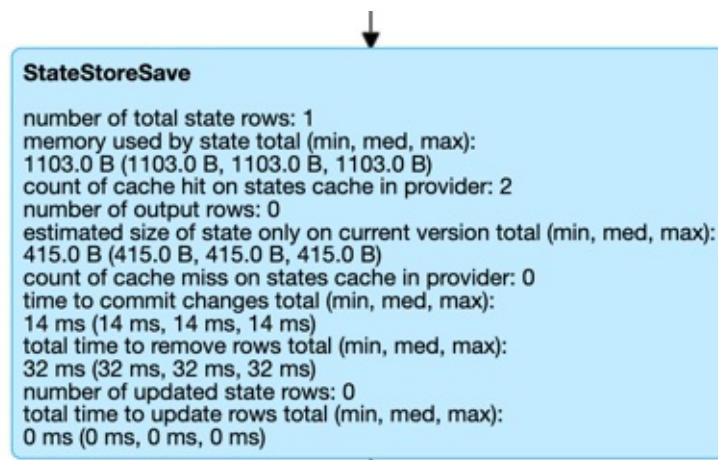


Figure 5. StateStoreSaveExec in web UI (Details for Query)

The following table shows how the performance metrics are computed (and so their exact meaning).

Name (in web UI)	Description
total time to update rows	<p>Time taken to read the input rows and store them in a state store (possibly filtering out expired rows per <code>watermarkPredicateForData</code> predicate)</p> <p>The number of rows stored is the <code>number of updated state rows</code> metric</p> <ul style="list-style-type: none"> <li>For <code>Append</code> output mode, the time taken to filter out expired rows (per the required <code>watermarkPredicateForData</code> predicate) and the <code>StreamingAggregationStateManager</code> to store rows in a state store</li> <li>For <code>Complete</code> output mode, the time taken to go over all the input rows and request the <code>StreamingAggregationStateManager</code> to store rows in a state store</li> <li>For <code>Update</code> output mode, the time taken to filter out expired rows (per the optional <code>watermarkPredicateForData</code> predicate) and the <code>StreamingAggregationStateManager</code> to store rows in a state store</li> </ul>
total time to remove rows	<ul style="list-style-type: none"> <li>For <code>Append</code> output mode, the time taken for the <code>StreamingAggregationStateManager</code> to remove all expired entries from a state store (per <code>watermarkPredicateForKeys</code> predicate) that is the total time of iterating over all entries in the state store (the number of entries removed from a state store is the difference between the number of output rows of the child operator and the <code>number of total state rows</code> metric)</li> <li>For <code>Complete</code> output mode, always 0</li> <li>For <code>Update</code> output mode, the time taken for the <code>StreamingAggregationStateManager</code> to remove all expired entries from a state store (per <code>watermarkPredicateForKeys</code> predicate)</li> </ul>
time to commit changes	<p>Time taken for the <code>StreamingAggregationStateManager</code> to commit changes to a state store</p> <ul style="list-style-type: none"> <li>For <code>Append</code> output mode, the metric does not seem to be used</li> <li>For <code>Complete</code> output mode, the number of rows in a StateStore (i.e. all values in a StateStore in the <code>StreamingAggregationStateManager</code> that should be</li> </ul>

number of output rows	<p>equivalent to the <a href="#">number of total state rows metric</a>)</p> <ul style="list-style-type: none"> <li>For <a href="#">Update</a> output mode, the number of rows that the <a href="#">StreamingAggregationStateManager</a> was requested to store in a state store (that did not expire per the optional <a href="#">watermarkPredicateForData</a> predicate) that is equivalent to the <a href="#">number of updated state rows metric</a>)</li> </ul>
number of total state rows	<p>Number of entries in a <a href="#">state store</a> at the very end of <a href="#">executing the StateStoreSaveExec operator</a> (aka <a href="#">numTotalStateRows</a>)</p> <p>Corresponds to <code> numRowsTotal </code> attribute in <code> stateOperators </code> in <a href="#">StreamingQueryProgress</a> (and is available as <code> sq.lastProgress.stateOperators </code> for an operator).</p>
number of updated state rows	<p>Number of the entries that <a href="#">were stored as updates in a state store</a> in a trigger and for the keys in the result rows of the upstream physical operator (aka <a href="#">numUpdatedStateRows</a>)</p> <ul style="list-style-type: none"> <li>For <a href="#">Append</a> output mode, the number of input rows that have not expired yet (per the required <a href="#">watermarkPredicateForData</a> predicate) and that the <a href="#">StreamingAggregationStateManager</a> was requested to store in a state store (the time taken is the <a href="#">total time to update rows</a> metric)</li> <li>For <a href="#">Complete</a> output mode, the number of input rows (which should be exactly the number of output rows from the <a href="#">child operator</a>)</li> <li>For <a href="#">Update</a> output mode, the number of rows that the <a href="#">StreamingAggregationStateManager</a> was requested to store in a state store (that did not expire per the optional <a href="#">watermarkPredicateForData</a> predicate) that is equivalent to the <a href="#">number of output rows metric</a>)</li> </ul> <p>Corresponds to <code> numRowsUpdated </code> attribute in <code> stateOperators </code> in <a href="#">StreamingQueryProgress</a> (and is available as <code> sq.lastProgress.stateOperators </code> for an operator).</p>
memory used by state	Estimated memory used by a <a href="#">StateStore</a> (aka <a href="#">stateMemory</a> ) after <code> stateStoreSaveExec </code> finished <a href="#">execution</a> (per the <a href="#">StateStoreMetrics</a> of the <a href="#">StateStore</a> )

## Creating StateStoreSaveExec Instance

`StateStoreSaveExec` takes the following to be created:

- **Key expressions**, i.e. Catalyst attributes for the grouping keys
- Execution-specific [StatefulOperatorStateInfo](#) (default: `None` )

- Execution-specific `output mode` (default: `None`)
- `Event-time watermark` (default: `None`)
- Version of the state format (based on the `spark.sql.streaming.aggregation.stateFormatVersion` configuration property)
- Child physical operator (`SparkPlan`)

## Executing Physical Operator (Generating RDD[InternalRow]) — `doExecute` Method

```
doExecute(): RDD[InternalRow]
```

**Note** `doExecute` is part of `SparkPlan` Contract to generate the runtime representation of an physical operator as a distributed computation over internal binary rows on Apache Spark (i.e. `RDD[InternalRow]` ).

Internally, `doExecute` initializes `metrics`.

**Note** `doExecute` requires that the optional `outputMode` is at this point defined (that should have happened when `IncrementalExecution` had prepared a streaming aggregation for execution).

`doExecute` executes `child` physical operator and creates a `StateStoreRDD` with `storeUpdateFunction` that:

1. Generates an unsafe projection to access the key field (using `keyExpressions` and the output schema of `child`).
2. Branches off per `output mode`: `Append`, `Complete` and `Update`.

`doExecute` throws an `UnsupportedOperationException` when executed with an invalid `output mode`:

```
Invalid output mode: [outputMode]
```

## Append Output Mode

**Note** `Append` is the default output mode when not specified explicitly.

**Note** `Append` output mode requires that a streaming query defines `event-time watermark` (e.g. using `withWatermark` operator) on the event-time column that is used in aggregation (directly or using `window` standard function).

For [Append](#) output mode, `doExecute` does the following:

1. Finds late (aggregate) rows from [child](#) physical operator (that have expired per [watermark](#))
2. Stores the late rows in the state store and increments the [numUpdatedStateRows](#) metric
3. Gets all the added (late) rows from the state store
4. Creates an iterator that removes the late rows from the state store when requested the next row and in the end commits the state updates

Tip

Refer to [Demo: Streaming Watermark with Aggregation in Append Output Mode](#) for an example of `stateStoreSaveExec` with [Append](#) output mode.

Caution

`FIXME` When is "Filtering state store on:" printed out?

1. Uses [watermarkPredicateForData](#) predicate to exclude matching rows and (like in [Complete](#) output mode) stores all the remaining rows in `StateStore`.
2. (like in [Complete](#) output mode) While storing the rows, increments [numUpdatedStateRows](#) metric (for every row) and records the total time in [allUpdatesTimeMs](#) metric.
3. Takes all the rows from `StateStore` and returns a `NextIterator` that:
  - In `getNext`, finds the first row that matches [watermarkPredicateForKeys](#) predicate, removes it from `StateStore`, and returns it back.  
If no row was found, `getNext` also marks the iterator as finished.
  - In `close`, records the time to iterate over all the rows in [allRemovalsTimeMs](#) metric, commits the updates to `Statestore` followed by recording the time in [commitTimeMs](#) metric and recording StateStore metrics.

## Complete Output Mode

For [Complete](#) output mode, `doExecute` does the following:

1. Takes all `unsafeRow` rows (from the parent iterator)
2. Stores the rows by key in the state store eagerly (i.e. all rows that are available in the parent iterator before proceeding)

3. Commits the state updates
4. In the end, reads the key-row pairs from the state store and passes the rows along (i.e. to the following physical operator)

The number of keys stored in the state store is recorded in `numUpdatedStateRows` metric.

**Note**

In `Complete` output mode the `numOutputRows` metric is exactly the `numTotalStateRows` metric.

**Tip**

Refer to [Demo: StateStoreSaveExec with Complete Output Mode](#) for an example of `StateStoreSaveExec` with `Complete` output mode.

- 
1. Stores all rows (as `UnsafeRow`) in `stateStore`.
  2. While storing the rows, increments `numUpdatedStateRows` metric (for every row) and records the total time in `allUpdatesTimeMs` metric.
  3. Records `0` in `allRemovalsTimeMs` metric.
  4. Commits the state updates to `statestore` and records the time in `commitTimeMs` metric.
  5. Records StateStore metrics.
  6. In the end, takes all the rows stored in `statestore` and increments `numOutputRows` metric.

## Update Output Mode

For `Update` output mode, `doExecute` returns an iterator that filters out late aggregate rows (per `watermark` if defined) and stores the "young" rows in the state store (one by one, i.e. every `next`).

With no more rows available, that removes the late rows from the state store (all at once) and commits the state updates.

**Tip**

Refer to [Demo: StateStoreSaveExec with Update Output Mode](#) for an example of `StateStoreSaveExec` with `update` output mode.

---

`doExecute` returns `Iterator` of rows that uses `watermarkPredicateForData` predicate to filter out late rows.

In `hasNext`, when rows are no longer available:

1. Records the total time to iterate over all the rows in `allUpdatesTimeMs` metric.
2. `removeKeysOlderThanWatermark` and records the time in `allRemovalsTimeMs` metric.
3. Commits the updates to `statestore` and records the time in `commitTimeMs` metric.
4. Records StateStore metrics.

In `next`, stores a row in `StateStore` and increments `numOutputRows` and `numUpdatedStateRows` metrics.

## Checking Out Whether Last Batch Execution Requires Another Non-Data Batch or Not

### — `shouldRunAnotherBatch` Method

```
shouldRunAnotherBatch(  
    newMetadata: OffsetSeqMetadata): Boolean
```

Note

`shouldRunAnotherBatch` is part of the [StateStoreWriter Contract](#) to indicate whether [MicroBatchExecution](#) should run another non-data batch (based on the updated [OffsetSeqMetadata](#) with the current event-time watermark and the batch timestamp).

`shouldRunAnotherBatch` is positive (`true`) when all of the following are met:

- [Output mode](#) is either [Append](#) or [Update](#)
- [Event-time watermark](#) is defined and is older (below) the current [event-time watermark](#) (of the given `offsetSeqMetadata`)

Otherwise, `shouldRunAnotherBatch` is negative (`false`).

# StreamingDeduplicateExec Unary Physical Operator for Streaming Deduplication

`StreamingDeduplicateExec` is a unary physical operator that writes state to `StateStore` with support for streaming watermark.

Note	A unary physical operator ( <code>UnaryExecNode</code> ) is a physical operator with a single child physical operator. Read up on <a href="#">UnaryExecNode</a> (and physical operators in general) in <a href="#">The Internals of Spark SQL book</a> .
------	---

`StreamingDeduplicateExec` is created exclusively when `StreamingDeduplicationStrategy` plans Deduplicate unary logical operators.

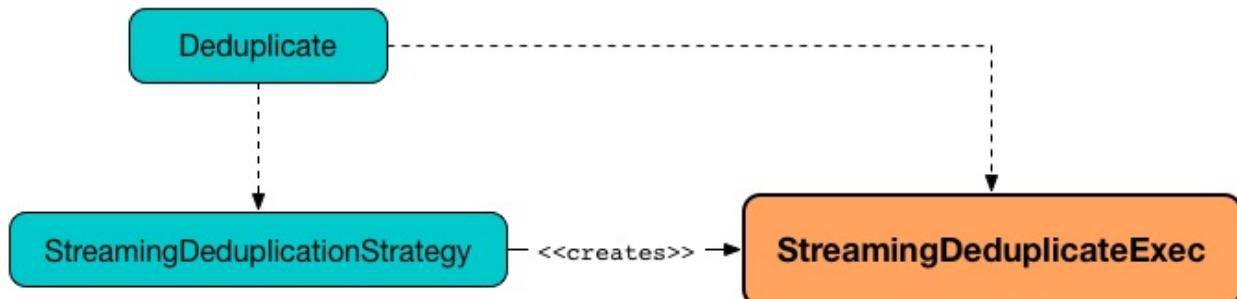


Figure 1. `StreamingDeduplicateExec` and `StreamingDeduplicationStrategy`

```

val uniqueValues = spark.
  readStream.
  format("rate").
  load.
  dropDuplicates("value") // <-- creates Deduplicate logical operator

scala> println(uniqueValues.queryExecution.logical.numberedTreeString)
00 Deduplicate [value#214L], true
01 +- StreamingRelation DataSource(org.apache.spark.sql.SparkSession@4785f176, rate, List(),
  None, List(), None, Map(), None), rate, [timestamp#213, value#214L]

scala> uniqueValues.explain
== Physical Plan ==
StreamingDeduplicate [value#214L], StatefulOperatorStateInfo(<unknown>, 5a65879c-67bc-4
e77-b417-6100db6a52a2, 0, 0), 0
+- Exchange hashpartitioning(value#214L, 200)
  +- StreamingRelation rate, [timestamp#213, value#214L]

// Start the query and hence StreamingDeduplicateExec
import scala.concurrent.duration._
import org.apache.spark.sql.streaming.{OutputMode, Trigger}
val sq = uniqueValues.
  
```

```
writeStream.  
  format("console").  
  option("truncate", false).  
  trigger(Trigger.ProcessingTime(10.seconds)).  
  outputMode(OutputMode.Update).  
  start  
  
// sorting not supported for non-aggregate queries  
// and so values are unsorted  
  
-----  
Batch: 0  
-----  
+---+---+  
|timestamp|value|  
+---+---+  
+---+---+  
  
-----  
Batch: 1  
-----  
+---+---+  
|timestamp | value |  
+---+---+  
| 2017-07-25 22:12:03.018 | 0 |  
| 2017-07-25 22:12:08.018 | 5 |  
| 2017-07-25 22:12:04.018 | 1 |  
| 2017-07-25 22:12:06.018 | 3 |  
| 2017-07-25 22:12:05.018 | 2 |  
| 2017-07-25 22:12:07.018 | 4 |  
+---+---+  
  
-----  
Batch: 2  
-----  
+---+---+  
|timestamp | value |  
+---+---+  
| 2017-07-25 22:12:10.018 | 7 |  
| 2017-07-25 22:12:09.018 | 6 |  
| 2017-07-25 22:12:12.018 | 9 |  
| 2017-07-25 22:12:13.018 | 10 |  
| 2017-07-25 22:12:15.018 | 12 |  
| 2017-07-25 22:12:11.018 | 8 |  
| 2017-07-25 22:12:14.018 | 11 |  
| 2017-07-25 22:12:16.018 | 13 |  
| 2017-07-25 22:12:17.018 | 14 |  
| 2017-07-25 22:12:18.018 | 15 |  
+---+---+  
  
// Eventually...  
sq.stop
```

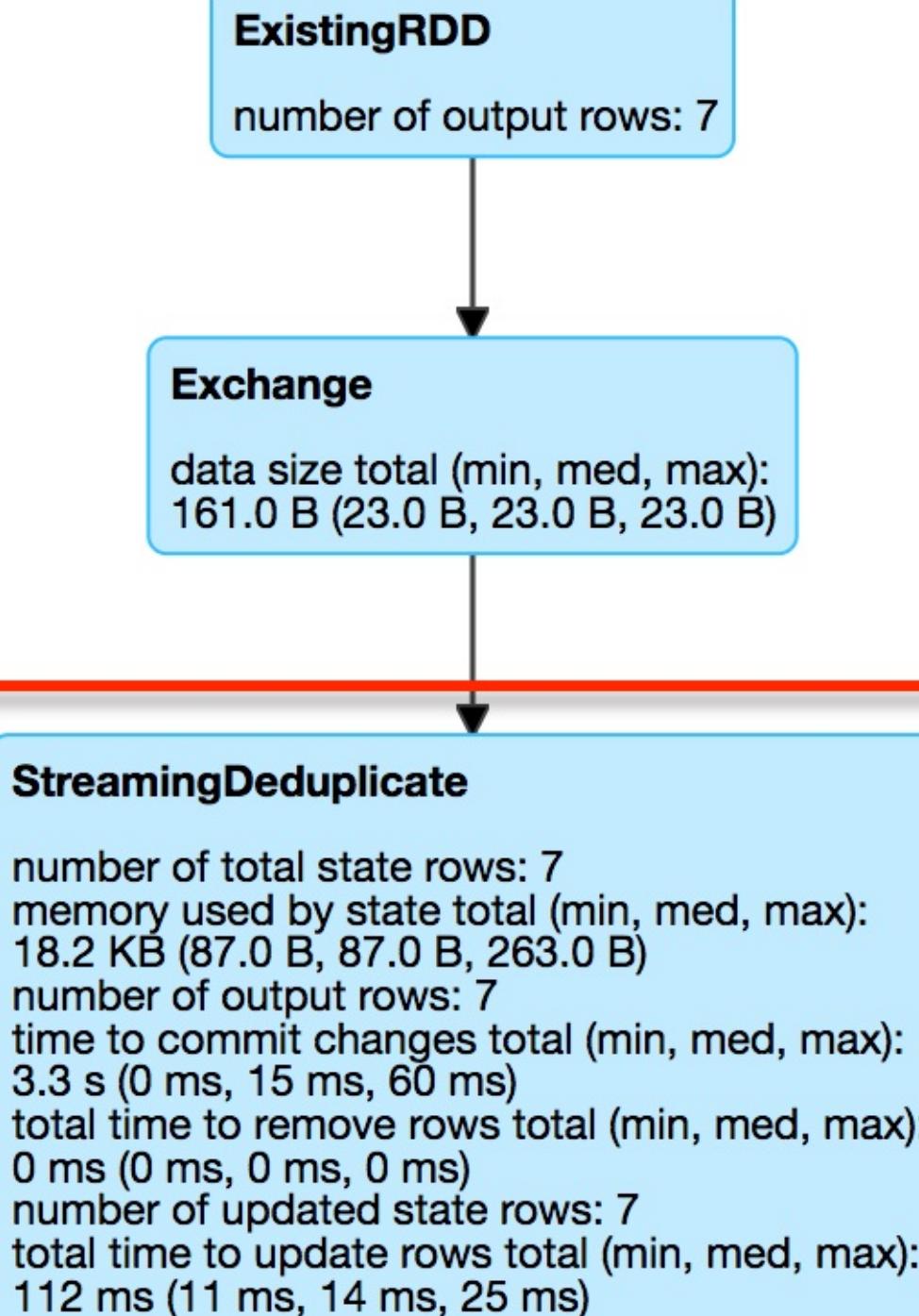


Figure 2. StreamingDeduplicateExec in web UI (Details for Query)

The output schema of `StreamingDeduplicateExec` is exactly the `child`'s output schema.

The output partitioning of `StreamingDeduplicateExec` is exactly the `child`'s output partitioning.

```
/*
// Start spark-shell with debugging and Kafka support
```

```

SPARK_SUBMIT_OPTS="-agentlib:jdwp=transport=dt_socket,server=y,suspend=n,address=500
5" \
./bin/spark-shell \
--packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.3.0-SNAPSHOT
*/
// Reading
val topic1 = spark.
readStream.
format("kafka").
option("subscribe", "topic1").
option("kafka.bootstrap.servers", "localhost:9092").
option("startingOffsets", "earliest").
load

// Processing with deduplication
// Don't use watermark
// The following won't work due to https://issues.apache.org/jira/browse/SPARK-21546
/**
val records = topic1.
  withColumn("eventtime", 'timestamp). // <-- just to put the right name given the pu
rpose
  withWatermark(eventTime = "eventtime", delayThreshold = "30 seconds"). // <-- use th
e renamed eventtime column
  dropDuplicates("value"). // dropDuplicates will use watermark
                            // only when eventTime column exists
  // include the watermark column => internal design leak?
  select('key cast "string", 'value cast "string", 'eventtime).
  as[(String, String, java.sql.Timestamp)]
*/

```

```

val records = topic1.
  dropDuplicates("value").
  select('key cast "string", 'value cast "string").
  as[(String, String)]

```

```

scala> records.explain
== Physical Plan ==
*Project [cast(key#0 as string) AS key#249, cast(value#1 as string) AS value#250]
+- StreamingDeduplicate [value#1], StatefulOperatorStateInfo(<unknown>,68198b93-6184-49
ae-8098-006c32cc6192,0,0), 0
  +- Exchange hashpartitioning(value#1, 200)
    +- *Project [key#0, value#1]
      +- StreamingRelation kafka, [key#0, value#1, topic#2, partition#3, offset#4L,
        timestamp#5, timestampType#6]

```

```

// Writing
import org.apache.spark.sql.streaming.{OutputMode, Trigger}
import scala.concurrent.duration.-
val sq = records.
writeStream.
format("console").
option("truncate", false).
trigger(Trigger.ProcessingTime(10.seconds)).

```

```
queryName("from-kafka-topic1-to-console").
outputMode(OutputMode.Update).
start

// Eventually...
sq.stop
```

**Tip**

Enable `INFO` logging level for `org.apache.spark.sql.execution.streaming.StreamingDeduplicateExec` to see what happens inside.

Add the following line to `conf/log4j.properties` :

```
log4j.logger.org.apache.spark.sql.execution.streaming.StreamingDeduplicateExec=I
```

Refer to [Logging](#).

## Executing Physical Operator (Generating RDD[InternalRow]) — `doExecute` Method

```
doExecute(): RDD[InternalRow]
```

**Note**

`doExecute` is part of `SparkPlan` Contract to generate the runtime representation of an physical operator as a distributed computation over internal binary rows on Apache Spark (i.e. `RDD[InternalRow]` ).

Internally, `doExecute` initializes [metrics](#).

`doExecute` executes `child` physical operator and creates a `StateStoreRDD` with `storeUpdateFunction` that:

1. Generates an unsafe projection to access the key field (using `keyExpressions` and the output schema of `child`).
2. Filters out rows from `Iterator[InternalRow]` that match `watermarkPredicateForData` (when defined and `timeoutConf` is `EventTimeTimeout` )
3. For every row (as `InternalRow`)
  - Extracts the key from the row (using the unsafe projection above)
  - Gets the saved state in `StateStore` for the key
  - (when there was a state for the key in the row) Filters out (aka *drops*) the row

- (when there was *no* state for the key in the row) Stores a new (and empty) state for the key and increments `numUpdatedStateRows` and `numOutputRows` metrics.
4. In the end, `storeUpdateFunction` creates a `CompletionIterator` that executes a completion function (aka `completionFunction`) after it has successfully iterated through all the elements (i.e. when a client has consumed all the rows).

The completion function does the following:

- Updates `allUpdatesTimeMs` metric (that is the total time to execute `storeUpdateFunction`)
- Updates `allRemovalsTimeMs` metric with the time taken to remove keys older than the watermark from the StateStore
- Updates `commitTimeMs` metric with the time taken to commit the changes to the StateStore
- Sets StateStore-specific metrics

## Creating StreamingDeduplicateExec Instance

`StreamingDeduplicateExec` takes the following when created:

- Duplicate keys (as used in `dropDuplicates` operator)
- Child physical operator (`SparkPlan`)
- `StatefulOperatorStateInfo`
- Event-time watermark

## Checking Out Whether Last Batch Execution Requires Another Non-Data Batch or Not — `shouldRunAnotherBatch` Method

```
shouldRunAnotherBatch(newMetadata: OffsetSeqMetadata): Boolean
```

Note	<code>shouldRunAnotherBatch</code> is part of the <code>StateStoreWriter Contract</code> to indicate whether <code>MicroBatchExecution</code> should run another non-data batch (based on the updated <code>OffsetSeqMetadata</code> with the current event-time watermark and the batch timestamp).
------	--

`shouldRunAnotherBatch` ...FIXME



# StreamingGlobalLimitExec Unary Physical Operator

`StreamingGlobalLimitExec` is a unary physical operator that represents a `Limit` logical operator of a streaming query at execution time.

Note

A unary physical operator (`UnaryExecNode`) is a physical operator with a single `child` physical operator.

Read up on [UnaryExecNode](#) (and physical operators in general) in [The Internals of Spark SQL](#) book.

`StreamingGlobalLimitExec` is created exclusively when [StreamingGlobalLimitStrategy](#) execution planning strategy is requested to plan a `Limit` logical operator (in the logical plan of a streaming query) for execution.

Note

`Limit` logical operator represents `Dataset.limit` operator in a logical query plan.

Read up on [Limit Logical Operator](#) in [The Internals of Spark SQL](#) book.

`StreamingGlobalLimitExec` is a [stateful physical operator](#) that can write to a state store.

`StreamingGlobalLimitExec` supports [Append](#) output mode only.

The optional properties, i.e. the [StatefulOperatorStateInfo](#) and the [output mode](#), are initially undefined when `StreamingGlobalLimitExec` is created. `StreamingGlobalLimitExec` is updated to hold execution-specific configuration when `IncrementalExecution` is requested to [prepare the logical plan \(of a streaming query\) for execution](#) (when the [state preparation rule](#) is executed).

## Creating StreamingGlobalLimitExec Instance

`StreamingGlobalLimitExec` takes the following to be created:

- **Streaming Limit**
- Child physical operator (`SparkPlan`)
- [StatefulOperatorStateInfo](#) (default: `None`)
- [OutputMode](#) (default: `None`)

`StreamingGlobalLimitExec` initializes the [internal properties](#).

## StreamingGlobalLimitExec as StateStoreWriter

StreamingGlobalLimitExec is a stateful physical operator that can write to a state store.

## Performance Metrics

StreamingGlobalLimitExec uses the performance metrics of the parent StateStoreWriter.

## Executing Physical Operator (Generating RDD[InternalRow]) — doExecute Method

```
doExecute(): RDD[InternalRow]
```

Note	doExecute is part of SparkPlan Contract to generate the runtime representation of an physical operator as a recipe for distributed computation over internal binary rows on Apache Spark ( RDD[InternalRow] ).
------	--

doExecute ...FIXME

## Internal Properties

Name	Description
keySchema	FIXME Used when...FIXME
valueSchema	FIXME Used when...FIXME

# StreamingRelationExec Leaf Physical Operator

StreamingRelationExec is a leaf physical operator (i.e. LeafExecNode) that...FIXME

StreamingRelationExec is created when StreamingRelationStrategy plans StreamingRelation and StreamingExecutionRelation logical operators.

```
scala> spark.version
res0: String = 2.3.0-SNAPSHOT

val rates = spark.
  readStream.
  format("rate").
  load

// StreamingRelation logical operator
scala> println(rates.queryExecution.logical.numberedTreeString)
00 StreamingRelation DataSource(org.apache.spark.sql.SparkSession@31ba0af0,rate,List(),
None,List(),None,Map(),None), rate, [timestamp#0, value#1L]

// StreamingRelationExec physical operator (shown without "Exec" suffix)
scala> rates.explain
== Physical Plan ==
StreamingRelation rate, [timestamp#0, value#1L]
```

StreamingRelationExec is not supposed to be executed and is used...FIXME

## Creating StreamingRelationExec Instance

StreamingRelationExec takes the following when created:

- The name of a streaming data source
- Output attributes

# StreamingSymmetricHashJoinExec Binary Physical Operator — Stream-Stream Joins

`StreamingSymmetricHashJoinExec` is a binary physical operator that represents a [stream-stream equi-join](#) at execution time.

Note

A binary physical operator (`BinaryExecNode`) is a physical operator with [left](#) and [right](#) child physical operators.  
Read up on [BinaryExecNode](#) (and physical operators in general) in [The Internals of Spark SQL](#) online book.

`StreamingSymmetricHashJoinExec` supports `Inner`, `LeftOuter`, and `RightOuter` join types (with the [left](#) and the [right](#) keys using the exact same data types).

`StreamingSymmetricHashJoinExec` is created exclusively when [StreamingJoinStrategy](#) execution planning strategy is requested to plan a logical query plan with a `Join` logical operator of two streaming queries with equality predicates (`EqualTo` and `EqualNullSafe`).

`StreamingSymmetricHashJoinExec` is given execution-specific configuration (i.e. [StatefulOperatorStateInfo](#), [event-time watermark](#), and [JoinStateWatermarkPredicates](#)) when [IncrementalExecution](#) is requested to plan a streaming query for execution (and uses the state preparation rule).

`StreamingSymmetricHashJoinExec` uses two [OneSideHashJoiners](#) (for the [left](#) and [right](#) sides of the join) to manage join state when processing partitions of the left and right sides of a stream-stream join.

`StreamingSymmetricHashJoinExec` is a [stateful physical operator](#) that writes to a state store.

## Creating StreamingSymmetricHashJoinExec Instance

`StreamingSymmetricHashJoinExec` takes the following to be created:

- Left keys (Catalyst expressions of the keys on the left side)
- Right keys (Catalyst expressions of the keys on the right side)
- [Join type](#)
- Join condition (`JoinConditionSplitPredicates`)
- [StatefulOperatorStateInfo](#)

- Event-Time Watermark
- Watermark Predicates for State Removal
- Physical operator on the left side ( `SparkPlan` )
- Physical operator on the right side ( `SparkPlan` )

`StreamingSymmetricHashJoinExec` initializes the [internal properties](#).

## Output Schema — `output` Method

```
output: Seq[Attribute]
```

Note	<code>output</code> is part of the <code>QueryPlan</code> Contract to describe the attributes of (the schema of) the output.
------	--

`output` schema depends on the [join type](#):

- For `Cross` and `Inner` (`InnerLike`) joins, it is the output schema of the `left` and `right` operators
- For `LeftOuter` joins, it is the output schema of the `left` operator with the attributes of the `right` operator with `nullability` flag enabled (`true`)
- For `RightOuter` joins, it is the output schema of the `right` operator with the attributes of the `left` operator with `nullability` flag enabled (`true`)

`output` throws an `IllegalArgumentException` for other join types:

```
[className] should not take [joinType] as the JoinType
```

## Output Partitioning — `outputPartitioning` Method

```
outputPartitioning: Partitioning
```

Note	<code>outputPartitioning</code> is part of the <code>SparkPlan</code> Contract to specify how data should be partitioned across different nodes in the cluster.
------	---

`outputPartitioning` depends on the [join type](#):

- For `Cross` and `Inner` (`InnerLike`) joins, it is a `PartitioningCollection` of the output partitioning of the `left` and `right` operators

- For `LeftOuter` joins, it is a `PartitioningCollection` of the output partitioning of the `left` operator
- For `RightOuter` joins, it is a `PartitioningCollection` of the output partitioning of the `right` operator

`outputPartitioning` throws an `IllegalArgumentException` for other join types:

`[className] should not take [joinType] as the JoinType`

## Event-Time Watermark — `eventTimeWatermark` Internal Property

`eventTimeWatermark: Option[Long]`

When `created`, `StreamingSymmetricHashJoinExec` can be given the `event-time watermark` of the current streaming micro-batch.

`eventTimeWatermark` is an optional property that is specified only after `IncrementalExecution` was requested to apply the `state preparation rule` to a physical query plan of a streaming query (to `optimize (prepare)` the physical plan of the streaming query once for `ContinuousExecution` and every trigger for `MicroBatchExecution` in their `queryPlanning` phases).

Note

- `eventTimeWatermark` is used when:
- `StreamingSymmetricHashJoinExec` is requested to `check out` whether the last batch execution requires another non-data batch or not
  - `OneSideHashJoiner` is requested to `storeAndJoinWithOtherSide`

## Watermark Predicates for State Removal — `stateWatermarkPredicates` Internal Property

`stateWatermarkPredicates: JoinStateWatermarkPredicates`

When `created`, `StreamingSymmetricHashJoinExec` is given a `JoinStateWatermarkPredicates` for the `left` and `right` join sides (using the `StreamingSymmetricHashJoinHelper` utility).

`stateWatermarkPredicates` contains the left and right predicates only when `IncrementalExecution` is requested to apply the `state preparation rule` to a physical query plan of a streaming query (to `optimize (prepare)` the physical plan of the streaming query

once for [ContinuousExecution](#) and every trigger for [MicroBatchExecution](#) in their **queryPlanning** phases).

<p><b>Note</b></p> <ul style="list-style-type: none"> <li>• <code>stateWatermarkPredicates</code> is used when <code>streamingSymmetricHashJoinExec</code> is requested for the following:</li> <li>• Process partitions of the left and right sides of the stream-stream join (and creating <a href="#">OneSideHashJoiners</a>)</li> <li>• Checking out whether the last batch execution requires another non-data batch or not</li> </ul>
---

## Required Partition Requirements — `requiredChildDistribution` Method

`requiredChildDistribution: Seq[Distribution]`

<p><b>Note</b></p> <p><code>requiredChildDistribution</code> is part of the <code>sparkPlan</code> Contract for the required partition requirements (aka <i>required child distribution</i>) of the input data, i.e. how the output of the children physical operators is split across partitions before this operator can be executed.</p> <p>Read up on <a href="#">SparkPlan Contract</a> in <a href="#">The Internals of Spark SQL</a> online book.</p>
---

`requiredChildDistribution` returns two `HashClusteredDistributions` for the `left` and `right` keys with the required [number of partitions](#) based on the [StatefulOperatorStateInfo](#).

<p><b>Note</b></p> <p><code>requiredChildDistribution</code> is used exclusively when <code>EnsureRequirements</code> physical query plan optimization is executed (and enforces partition requirements).</p> <p>Read up on <a href="#">EnsureRequirements Physical Query Optimization</a> in <a href="#">The Internals of Spark SQL</a> online book.</p>
---

<p><b>Note</b></p> <p><code>HashClusteredDistribution</code> becomes <code>HashPartitioning</code> at execution that distributes rows across partitions (generates partition IDs of rows) based on <code>Murmur3Hash</code> of the join expressions (separately for the <code>left</code> and <code>right</code> keys) modulo the required number of partitions.</p> <p>Read up on <a href="#">HashClusteredDistribution</a> in <a href="#">The Internals of Spark SQL</a> online book.</p>
---

## Performance Metrics (SQLMetrics)

`StreamingSymmetricHashJoinExec` uses the performance metrics as other stateful physical operators that write to a state store.

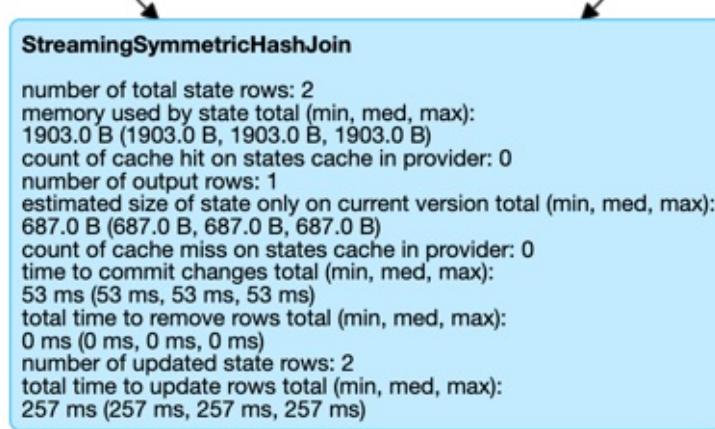


Figure 1. `StreamingSymmetricHashJoinExec` in web UI (Details for Query)

The following table shows how the performance metrics are computed (and so their exact meaning).

Name (in web UI)	Description
total time to update rows	Processing time of all rows
total time to remove rows	
time to commit changes	
number of output rows	Total number of output rows
number of total state rows	
number of updated state rows	<a href="#">Number of updated state rows</a> of the <a href="#">left</a> and <a href="#">right</a> OneSideHashJoiners
memory used by state	

## Checking Out Whether Last Batch Execution Requires Another Non-Data Batch or Not — `shouldRunAnotherBatch` Method

```
shouldRunAnotherBatch(  
    newMetadata: OffsetSeqMetadata): Boolean
```

**Note**

`shouldRunAnotherBatch` is part of the [StateStoreWriter Contract](#) to indicate whether [MicroBatchExecution](#) should run another non-data batch (based on the updated [OffsetSeqMetadata](#) with the current event-time watermark and the batch timestamp).

`shouldRunAnotherBatch` is positive (`true`) when all of the following are positive:

- Either the [left](#) or [right](#) join state watermark predicates are defined (in the [JoinStateWatermarkPredicates](#))
- [Event-time watermark](#) threshold (of the [StreamingSymmetricHashJoinExec](#) operator) is defined and the current [event-time watermark](#) threshold of the given [OffsetSeqMetadata](#) is above (*greater than*) it, i.e. moved above

`shouldRunAnotherBatch` is negative (`false`) otherwise.

## Executing Physical Operator (Generating RDD[InternalRow]) — `doExecute` Method

```
doExecute(): RDD[InternalRow]
```

**Note**

`doExecute` is part of [SparkPlan Contract](#) to generate the runtime representation of a physical operator as a recipe for distributed computation over internal binary rows on Apache Spark (`RDD[InternalRow]`).

`doExecute` first requests the [StreamingQueryManager](#) for the [StateStoreCoordinatorRef](#) to the [StateStoreCoordinator](#) RPC endpoint (for the driver).

`doExecute` then uses [SymmetricHashJoinStateManager](#) utility to get the names of the state stores for the [left](#) and [right](#) sides of the streaming join.

In the end, `doExecute` requests the [left](#) and [right](#) child physical operators to execute (generate an RDD) and then [stateStoreAwareZipPartitions](#) with [processPartitions](#) (and with the [StateStoreCoordinatorRef](#) and the state stores).

## Processing Partitions of Left and Right Sides of Stream-Stream Join — `processPartitions` Internal Method

```
processPartitions(  
    leftInputIter: Iterator[InternalRow],  
    rightInputIter: Iterator[InternalRow]): Iterator[InternalRow]
```

`processPartitions` records the current time (as `updateStartTimeNs` for the total time to update rows performance metric in `onOutputCompletion`).

`processPartitions` creates a new predicate (`postJoinFilter`) based on the `bothSides` of the `JoinConditionSplitPredicates` if defined or `true` literal.

`processPartitions` creates a `OneSideHashJoiner` for the `LeftSide` and all other properties for the left-hand join side (`leftSideJoiner`).

`processPartitions` creates a `OneSideHashJoiner` for the `RightSide` and all other properties for the right-hand join side (`rightSideJoiner`).

`processPartitions` requests the `OneSideHashJoiner` for the left-hand join side to `storeAndJoinWithOtherSide` with the right-hand side one (that creates a `leftOutputIter` row iterator) and the `OneSideHashJoiner` for the right-hand join side to do the same with the left-hand side one (and creates a `rightOutputIter` row iterator).

`processPartitions` records the current time (as `innerOutputCompletionTimeNs` for the total time to remove rows performance metric in `onOutputCompletion`).

`processPartitions` creates a `CompletionIterator` with the left and right output iterators (with the rows of the `leftOutputIter` first followed by `rightOutputIter`). When no rows are left to process, the `CompletionIterator` records the completion time.

`processPartitions` creates a join-specific output `Iterator[InternalRow]` of the output rows based on the `join type` (of the `StreamingSymmetricHashJoinExec`):

- For `Inner` joins, `processPartitions` simply uses the `output iterator of the left and right rows`
- For `LeftOuter` joins, `processPartitions` ...
- For `RightOuter` joins, `processPartitions` ...
- For other joins, `processPartitions` simply throws an `IllegalArgumentException`.

`processPartitions` creates an `UnsafeProjection` for the `output` (and the output of the `left` and `right` child operators) that counts all the rows of the `join-specific output iterator` (as the `numOutputRows` metric) and generate an output projection.

In the end, `processPartitions` returns a `CompletionIterator` with the `output iterator with the rows counted (as numOutputRows metric)` and `onOutputCompletion` completion function.

Note	<code>processPartitions</code> is used exclusively when <code>StreamingSymmetricHashJoinExec</code> physical operator is requested to <code>execute</code> .
------	--

## Calculating Performance Metrics (Output Completion Callback) — `onOutputCompletion` Internal Method

`onOutputCompletion: Unit`

`onOutputCompletion` calculates the [total time to update rows](#) performance metric (that is the time since the [processPartitions](#) was executed).

`onOutputCompletion` adds the time for the inner join to complete (since [innerOutputCompletionTimeNs](#) time marker) to the [total time to remove rows](#) performance metric.

`onOutputCompletion` records the time to [remove old state](#) (per the [join state watermark predicate](#) for the [left](#) and the [right](#) streaming queries) and adds it to the [total time to remove rows](#) performance metric.

**Note**

`onOutputCompletion` triggers the [old state removal](#) eagerly by iterating over the state rows to be deleted.

`onOutputCompletion` records the time for the [left](#) and [right](#) [OneSideHashJoiners](#) to [commit any state changes](#) that becomes the [time to commit changes](#) performance metric.

`onOutputCompletion` calculates the [number of updated state rows](#) performance metric (as the [number of updated state rows](#) of the [left](#) and [right](#) streaming queries).

`onOutputCompletion` calculates the [number of total state rows](#) performance metric (as the sum of the [number of keys](#) in the [KeyWithIndexToValueStore](#) of the [left](#) and [right](#) streaming queries).

`onOutputCompletion` calculates the [memory used by state](#) performance metric (as the sum of the [memory used by the KeyToNumValuesStore](#) and [KeyWithIndexToValueStore](#) of the [left](#) and [right](#) streams).

In the end, `onOutputCompletion` calculates the [custom metrics](#).

## Internal Properties

Name	Description
hadoopConfBcast	Hadoop Configuration broadcast (to the Spark cluster) Used exclusively to <a href="#">create a SymmetricHashJoinStateManager</a>
joinStateManager	<a href="#">SymmetricHashJoinStateManager</a> Used when <code>OneSideHashJoiner</code> is requested to <code>storeAndJoinWithOtherSide</code> , <code>removeOldState</code> , <code>commitStateAndGetMetrics</code> , and for the <code>values</code> for a given key
nullLeft	<code>GenericInternalRow</code> of the size of the output schema of the <a href="#">left physical operator</a>
nullRight	<code>GenericInternalRow</code> of the size of the output schema of the <a href="#">right physical operator</a>
storeConf	<a href="#">StateStoreConf</a> Used exclusively to <a href="#">create a SymmetricHashJoinStateManager</a>

# FlatMapGroupsWithStateStrategy Execution Planning Strategy for FlatMapGroupsWithState Logical Operator

`FlatMapGroupsWithStateStrategy` is an execution planning strategy that can plan streaming queries with `FlatMapGroupsWithState` unary logical operators to `FlatMapGroupsWithStateExec` physical operator (with undefined `StatefulOperatorStateInfo`, `batchTimestampMs`, and `eventTimeWatermark`).

**Tip** Read up on [Execution Planning Strategies](#) in [The Internals of Spark SQL](#) book.

`FlatMapGroupsWithStateStrategy` is used exclusively when `IncrementalExecution` is requested to plan a streaming query.

## Demo: Using FlatMapGroupsWithStateStrategy

```

import org.apache.spark.sql.streaming.GroupState
val stateFunc = (key: Long, values: Iterator[(Timestamp, Long)], state: GroupState[Long]) => {
  Iterator((key, values.size))
}
import java.sql.Timestamp
import org.apache.spark.sql.streaming.{GroupStateTimeout, OutputMode}
val numGroups = spark.
  readStream.
  format("rate").
  load.
  as[(Timestamp, Long)].
  groupByKey { case (time, value) => value % 2 }.
  flatMapGroupsWithState(OutputMode.Update, GroupStateTimeout.NoTimeout)(stateFunc)

scala> numGroups.explain(true)
== Parsed Logical Plan ==
'SerializeFromObject [assertnonnull(assertnonnull(input[0, scala.Tuple2, true]))._1 AS
 _1#267L, assertnonnull(assertnonnull(input[0, scala.Tuple2, true]))._2 AS _2#268]
+- 'FlatMapGroupsWithState <function3>, unresolveddeserializer(upcast(getcolumnbyordin
 al(0, LongType), LongType, - root class: "scala.Long"), value#262L), unresolveddeseria
 lizer(newInstance(class scala.Tuple2), timestamp#253, value#254L), [value#262L], [time
 stamp#253, value#254L], obj#266: scala.Tuple2, class[value[0]: bigint], Update, false,
 NoTimeout
   +- AppendColumns <function1>, class scala.Tuple2, [StructField(_1, TimestampType, tru
 e), StructField(_2, LongType, false)], newInstance(class scala.Tuple2), [input[0, bigint
 , false] AS value#262L]
     +- StreamingRelation DataSource(org.apache.spark.sql.SparkSession@38bcac50, rate,
 List(), None, List(), None, Map(), None), rate, [timestamp#253, value#254L]

...
== Physical Plan ==
*SerializeFromObject [assertnonnull(input[0, scala.Tuple2, true])._1 AS _1#267L, asse
 tnonnull(input[0, scala.Tuple2, true])._2 AS _2#268]
+- FlatMapGroupsWithState <function3>, value#262: bigint, newInstance(class scala.Tupl
 e2), [value#262L], [timestamp#253, value#254L], obj#266: scala.Tuple2, StatefulOperato
 rStateInfo(<unknown>, 84b5dccb-3fa6-4343-a99c-6fa5490c9b33, 0, 0), class[value[0]: bigi
 nt], Update, NoTimeout, 0, 0
  +- *Sort [value#262L ASC NULLS FIRST], false, 0
    +- Exchange hashpartitioning(value#262L, 200)
      +- AppendColumns <function1>, newInstance(class scala.Tuple2), [input[0, bigi
 nt, false] AS value#262L]
        +- StreamingRelation rate, [timestamp#253, value#254L]

```

# StatefulAggregationStrategy Execution Planning Strategy — EventTimeWatermark and Aggregate Logical Operators

`StatefulAggregationStrategy` is an execution planning strategy that is used to plan streaming queries with the two logical operators:

- `EventTimeWatermark` logical operator (for `Dataset.withWatermark` operator)
- `Aggregate` logical operator (for `Dataset.groupBy` and `Dataset.groupByKey` operators, and `GROUP BY` SQL clause)

**Tip** Read up on [Execution Planning Strategies](#) in [The Internals of Spark SQL](#) book.

`StatefulAggregationStrategy` is used exclusively when `IncrementalExecution` is requested to plan a streaming query.

`StatefulAggregationStrategy` is available using `SessionState`.

```
spark.sessionState.planner.StatefulAggregationStrategy
```

Table 1. StatefulAggregationStrategy's Logical to Physical Operator Conversions

Logical Operator	Physical Operator
<code>EventTimeWatermark</code>	<code>EventTimeWatermarkExec</code>
<code>Aggregate</code>	<p>In the order of preference:</p> <ol style="list-style-type: none"> <li>1. <code>HashAggregateExec</code></li> <li>2. <code>ObjectHashAggregateExec</code></li> <li>3. <code>SortAggregateExec</code></li> </ol>
	<p><b>Tip</b> Read up on <a href="#">Aggregation Execution Planning Strategy for Aggregate Physical Operators</a> in <a href="#">The Internals of Spark SQL</a> book.</p>

```

val counts = spark.
  readStream.
  format("rate").
  load.
  groupBy(window($"timestamp", "5 seconds") as "group").
  agg(count("value") as "count").
  orderBy("group")
scala> counts.explain
== Physical Plan ==
*Sort [group#6 ASC NULLS FIRST], true, 0
+- Exchange rangepartitioning(group#6 ASC NULLS FIRST, 200)
  +- *HashAggregate(keys=[window#13], functions=[count(value#1L)])
    +- StateStoreSave [window#13], StatefulOperatorStateInfo(<unknown>, 736d67c2-6daa
-4c4c-9c4b-c12b15af20f4, 0, 0), Append, 0
    +- *HashAggregate(keys=[window#13], functions=[merge_count(value#1L)])
      +- StateStoreRestore [window#13], StatefulOperatorStateInfo(<unknown>, 736d
67c2-6daa-4c4c-9c4b-c12b15af20f4, 0, 0)
      +- *HashAggregate(keys=[window#13], functions=[merge_count(value#1L)])
        +- Exchange hashpartitioning(window#13, 200)
          +- *HashAggregate(keys=[window#13], functions=[partial_count(valu
e#1L)])
            +- *Project [named_struct(start, precisetimestampconversion(((CASE WHEN (cast(CEIL((cast((precisetimestampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0)) as double) = (cast((precisetimestampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0)) THEN (CEIL((cast((precisetimestampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0)) + 1) ELSE CEIL((cast((precisetimestampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0)) END + 0) - 1) * 5000000) + 0), LongType, Times
tampType), end, precisetimestampconversion((((CASE WHEN (cast(CEIL((cast((precisetime
stampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0)) as
double) = (cast((precisetimestampconversion(timestamp#0, TimestampType, LongType) - 0
) as double) / 5000000.0)) THEN (CEIL((cast((precisetimestampconversion(timestamp#0, T
imestampType, LongType) - 0) as double) / 5000000.0)) + 1) ELSE CEIL((cast((precisetim
estampconversion(timestamp#0, TimestampType, LongType) - 0) as double) / 5000000.0)) E
ND + 0) - 1) * 5000000) + 5000000), LongType, TimestampType)) AS window#13, value#1L]
            +- *Filter isnullobject(timestamp#0)
              +- StreamingRelation rate, [timestamp#0, value#1L]

import org.apache.spark.sql.streaming.{OutputMode, Trigger}
import scala.concurrent.duration.-
val consoleOutput = counts.
  writeStream.
  format("console").
  option("truncate", false).
  trigger(Trigger.ProcessingTime(10.seconds)).
  queryName("counts").
  outputMode(OutputMode.Complete). // <-- required for groupBy
  start

// Eventually...
consoleOutput.stop

```

## Selecting Aggregate Physical Operator Given Aggregate Expressions — `AggUtils.planStreamingAggregation` Internal Method

```
planStreamingAggregation(
    groupingExpressions: Seq[NamedExpression],
    functionsWithoutDistinct: Seq[AggregateExpression],
    resultExpressions: Seq[NamedExpression],
    child: SparkPlan): Seq[SparkPlan]
```

`planStreamingAggregation` takes the grouping attributes (from `groupingExpressions` ).

Note	<code>groupingExpressions</code> corresponds to the grouping function in <a href="#">groupBy</a> operator.
------	--

`planStreamingAggregation` creates an aggregate physical operator (called `partialAggregate` ) with:

- `requiredChildDistributionExpressions` `undefined` (i.e. `None` )
- `initialInputBufferOffset` `as 0`
- `functionsWithoutDistinct` in `Partial` mode
- `child` operator as the input `child`

Note	<p><code>planStreamingAggregation</code> creates one of the following aggregate physical operators (in the order of preference):</p> <ol style="list-style-type: none"> <li>1. <code>HashAggregateExec</code></li> <li>2. <code>ObjectHashAggregateExec</code></li> <li>3. <code>SortAggregateExec</code></li> </ol> <p><code>planStreamingAggregation</code> uses <code>AggUtils.createAggregate</code> method to select an aggregate physical operator that you can read about in <a href="#">Selecting Aggregate Physical Operator Given Aggregate Expressions — <code>AggUtils.createAggregate</code> Internal Method</a> in <a href="#">Mastering Apache Spark 2</a> gitbook.</p>
------	--

`planStreamingAggregation` creates an aggregate physical operator (called `partialMerged1` ) with:

- `requiredChildDistributionExpressions` based on the input `groupingExpressions`
- `initialInputBufferOffset` as the length of `groupingExpressions`
- `functionsWithoutDistinct` in `PartialMerge` mode
- `child` operator as [partialAggregate](#) aggregate physical operator created above

`planStreamingAggregation` creates `StateStoreRestoreExec` with the grouping attributes, `statefulOperatorStateInfo`, and `partialMerged1` aggregate physical operator created above.

`planStreamingAggregation` creates an aggregate physical operator (called `partialMerged2`) with:

- `child` operator as `StateStoreRestoreExec` physical operator created above

Note	The only difference between <code>partialMerged1</code> and <code>partialMerged2</code> steps is the child physical operator.
------	---

`planStreamingAggregation` creates `StateStoreSaveExec` with:

- the grouping attributes based on the input `groupingExpressions`
- No `stateInfo`, `outputMode` and `eventTimeWatermark`
- `child` operator as `partialMerged2` aggregate physical operator created above

In the end, `planStreamingAggregation` creates the final aggregate physical operator (called `finalAndCompleteAggregate`) with:

- `requiredChildDistributionExpressions` based on the input `groupingExpressions`
- `initialInputBufferOffset` as the length of `groupingExpressions`
- `functionsWithoutDistinct` in `Final` mode
- `child` operator as `StateStoreSaveExec` physical operator created above

Note	<code>planStreamingAggregation</code> is used exclusively when <code>StatefulAggregationStrategy</code> plans a streaming aggregation.
------	--

# StreamingDeduplicationStrategy Execution Planning Strategy for Deduplicate Logical Operator

`StreamingDeduplicationStrategy` is an execution planning strategy that can plan streaming queries with `Deduplicate` logical operators (over streaming queries) to `StreamingDeduplicateExec` physical operators.

**Tip** Read up on [Execution Planning Strategies](#) in [The Internals of Spark SQL](#) book.

**Note** `Deduplicate` logical operator represents `Dataset.dropDuplicates` operator in a logical query plan.

`StreamingDeduplicationStrategy` is available using `SessionState`.

```
spark.sessionState.planner.StreamingDeduplicationStrategy
```

## Demo: Using StreamingDeduplicationStrategy

FIXME

# StreamingGlobalLimitStrategy Execution Planning Strategy

`StreamingGlobalLimitStrategy` is an execution planning strategy that can plan streaming queries with `ReturnAnswer` and `Limit` logical operators (over streaming queries) with the `Append` output mode to `StreamingGlobalLimitExec` physical operator.

Tip

Read up on [Execution Planning Strategies](#) in [The Internals of Spark SQL](#) book.

`StreamingGlobalLimitStrategy` is used (and [created](#)) exclusively when `IncrementalExecution` is requested to plan a streaming query.

`StreamingGlobalLimitStrategy` takes a single `OutputMode` to be created (which is the `OutputMode` of the `IncrementalExecution`).

## Demo: Using StreamingGlobalLimitStrategy

FIXME

# StreamingJoinStrategy Execution Planning Strategy — Stream-Stream Equi-Joins

`StreamingJoinStrategy` is an execution planning strategy that can plan streaming queries with `Join` logical operators of two streaming queries to a `StreamingSymmetricHashJoinExec` physical operator.

Tip

Read up on [Execution Planning Strategies](#) in [The Internals of Spark SQL](#) online book.

`StreamingJoinStrategy` throws an `AnalysisException` when applied to a `Join` logical operator with no equality predicate:

Stream-stream join without equality predicate is not supported

`StreamingJoinStrategy` is used exclusively when [IncrementalExecution](#) is requested to plan a streaming query.

Tip

`StreamingJoinStrategy` does not print out any messages to the logs. `StreamingJoinStrategy` however uses `ExtractEquiJoinKeys` Scala extractor for destructuring `Join` logical operators that does print out DEBUG messages to the logs.

Read up on [ExtractEquiJoinKeys](#) in [The Internals of Spark SQL](#) online book.

Enable `ALL` logging level for

`org.apache.spark.sql.catalyst.planning.ExtractEquiJoinKeys` to see what happens inside.

Add the following line to `conf/log4j.properties` :

```
log4j.logger.org.apache.spark.sql.catalyst.planning.ExtractEquiJoinKeys=ALL
```

Refer to [Logging](#).

# StreamingRelationStrategy Execution Planning Strategy for StreamingRelation and StreamingExecutionRelation Logical Operators

`StreamingRelationStrategy` is an execution planning strategy that can plan streaming queries with `StreamingRelation`, `StreamingExecutionRelation`, and `StreamingRelationV2` logical operators to `StreamingRelationExec` physical operators.

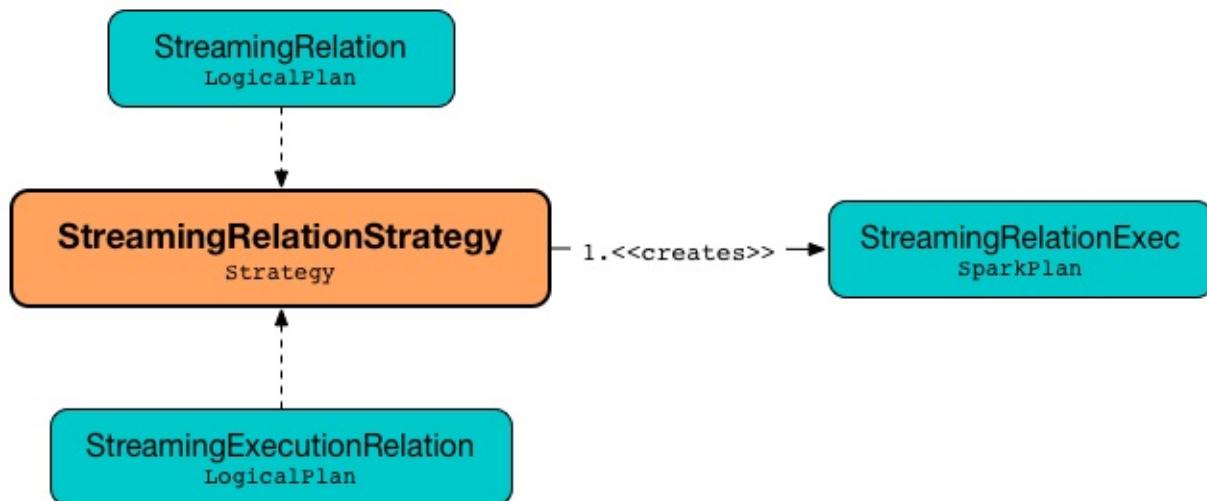


Figure 1. StreamingRelationStrategy, StreamingRelation, StreamingExecutionRelation and StreamingRelationExec Operators

Tip	Read up on <a href="#">Execution Planning Strategies</a> in <a href="#">The Internals of Spark SQL</a> book.
-----	--

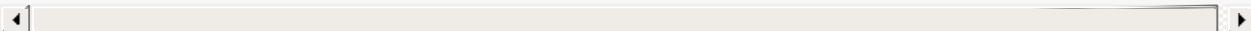
`StreamingRelationStrategy` is used exclusively when `IncrementalExecution` is requested to plan a streaming query.

`StreamingRelationStrategy` is available using `SessionState` (of a `SparkSession` ).

```
spark.sessionState.planner.StreamingRelationStrategy
```

## Demo: Using StreamingRelationStrategy

```
val rates = spark.  
  readStream.  
  format("rate").  
  load // <-- gives a streaming Dataset with a logical plan with StreamingRelation logical operator  
  
// StreamingRelation logical operator for the rate streaming source  
scala> println(rates.queryExecution.logical.numberedTreeString)  
00 StreamingRelation DataSource(org.apache.spark.sql.SparkSession@31ba0af0, rate, List(),  
None, List(), None, Map(), None), rate, [timestamp#0, value#1L]  
  
// StreamingRelationExec physical operator (shown without "Exec" suffix)  
scala> rates.explain  
== Physical Plan ==  
StreamingRelation rate, [timestamp#0, value#1L]  
  
// Let's do the planning manually  
import spark.sessionState.planner.StreamingRelationStrategy  
val physicalPlan = StreamingRelationStrategy.apply(rates.queryExecution.logical).head  
scala> println(physicalPlan.numberedTreeString)  
00 StreamingRelation rate, [timestamp#0, value#1L]
```



# UnsupportedOperationChecker

`UnsupportedOperationChecker` checks whether the logical plan of a streaming query uses supported operations only.

Note	<code>UnsupportedOperationChecker</code> is used exclusively when the internal <code>spark.sql.streaming.unsupportedOperationCheck</code> Spark property is enabled (which is by default).
------	--

Note	<code>UnsupportedOperationChecker</code> comes actually with two methods, i.e. <code>checkForBatch</code> and <code>checkForStreaming</code> , whose names reveal the different flavours of Spark SQL (as of 2.0), i.e. batch and streaming, respectively.
------	--

The Spark Structured Streaming gitbook is solely focused on `checkForStreaming` method.

## checkForStreaming Method

```
checkForStreaming(  
    plan: LogicalPlan,  
    outputMode: OutputMode): Unit
```

`checkForStreaming` asserts that the following requirements hold:

1. Only one streaming aggregation is allowed
2. Streaming aggregation with Append output mode requires watermark (on the grouping expressions)
3. Multiple flatMapGroupsWithState operators are only allowed with Append output mode

`checkForStreaming ...FIXME`

`checkForStreaming` finds all streaming aggregates (i.e. `Aggregate` logical operators with streaming sources).

Note	<code>Aggregate</code> logical operator represents <code>Dataset.groupBy</code> and <code>Dataset.groupByKey</code> operators (and SQL's <code>GROUP BY</code> clause) in a logical query plan.
------	---

`checkForStreaming` asserts that there is exactly one streaming aggregation in a streaming query.

Otherwise, `checkForStreaming` reports a `AnalysisException`:

Multiple streaming aggregations are not supported with streaming DataFrames/Datasets

`checkForStreaming` asserts that `watermark` was defined for a streaming aggregation with `Append` output mode (on at least one of the grouping expressions).

Otherwise, `checkForStreaming` reports a `AnalysisException`:

Append output mode not supported when there are streaming aggregations on streaming DataFrames/DataSets without watermark

Caution	FIXME
---------	-------

`checkForStreaming` counts all `FlatMapGroupsWithState` logical operators (on streaming Datasets with `isMapGroupsWithState` flag disabled).

Note	<code>FlatMapGroupsWithState</code> logical operator represents <code>KeyValueGroupedDataset.mapGroupsWithState</code> and <code>KeyValueGroupedDataset.flatMapGroupsWithState</code> operators in a logical query plan.
Note	<code>FlatMapGroupsWithState.isMapGroupsWithState</code> flag is disabled when... FIXME

`checkForStreaming` asserts that multiple `FlatMapGroupsWithState` logical operators are only used when:

- `outputMode` is `Append` output mode
- `outputMode` of the `FlatMapGroupsWithState` logical operators is also `Append` output mode

Caution	FIXME Reference to an example in <code>flatMapGroupsWithState</code>
---------	--

Otherwise, `checkForStreaming` reports a `AnalysisException`:

Multiple `flatMapGroupsWithStates` are not supported when they are not all in append mode or the output mode is not append on a streaming DataFrames/Datasets

Caution	FIXME
---------	-------

Note	<p><code>checkForStreaming</code> is used exclusively when <code>StreamingQueryManager</code> is requested to <a href="#">create a <code>StreamingQueryWrapper</code></a> (for starting a streaming query), but only when the internal <code>spark.sql.streaming.unsupportedOperationCheck</code> Spark property is enabled (which is by default).</p>
------	--