

Spring Cloud Data Flow Server for Mesos

1.0.0.BUILD-SNAPSHOT

Copyright © 2013-2016 Pivotal Software, Inc.

Copies of this document may be made for your own use and for distribution to others, provided that you do not charge any fee for such copies and further provided that each copy contains this Copyright Notice, whether distributed in print or electronically.

Table of Contents

I. Introduction	1
1. Introducing Spring Cloud Data Flow for Mesos and Marathon	2
2. Spring Cloud Data Flow	3
3. Spring Cloud Stream	4
4. Spring Cloud Task	5
II. Architecture	6
5. Introduction	7
6. Microservice Architectural Style	9
6.1. Comparison to other Platform architectures	9
7. Streaming Applications	11
7.1. Imperative Programming Model	11
7.2. Functional Programming Model	11
8. Streams	12
8.1. Topologies	12
8.2. Concurrency	12
8.3. Partitioning	12
8.4. Message Delivery Guarantees	13
9. Analytics	14
10. Task Applications	15
11. Data Flow Server	16
11.1. Endpoints	16
11.2. Customization	16
11.3. Security	17
12. Runtime	18
12.1. Fault Tolerance	18
12.2. Resource Management	18
12.3. Scaling at runtime	18
12.4. Application Versioning	18
III. Getting Started	19
13. Deploying Streams and Tasks on Mesos and Marathon/Chronos	20
IV. Streams	26
14. Introduction	27
15. Stream DSL	28
16. Register a Stream App	29
16.1. Whitelisting application properties	30
17. Creating a Stream	32
17.1. Application properties	32
Passing application properties when creating a stream	32
Passing application properties when deploying a stream	34
Passing stream partition properties during stream deployment	34
Overriding application properties during stream deployment	35
17.2. Deployment properties	35
Passing instance count as deployment property	35
Inline vs file reference properties	35
18. Destroying a Stream	37
19. Deploying and Undeploying Streams	38
20. Other Source and Sink Application Types	39

21. Simple Stream Processing	40
22. Stateful Stream Processing	41
23. Tap a Stream	42
24. Using Labels in a Stream	43
25. Explicit Broker Destinations in a Stream	44
26. Directed Graphs in a Stream	45
26.1. Common application properties	45
27. Stream applications with multiple binder configurations	46
V. Dashboard	47
28. Introduction	48
29. Apps	49
30. Runtime	50
31. Streams	51
32. Create Stream	52
33. Tasks	53
33.1. Apps	53
Create a Task Definition from a selected Task App	53
View Task App Details	54
33.2. Definitions	54
Launching Tasks	54
33.3. Executions	54
34. Jobs	55
34.1. List job executions	55
Job execution details	56
Step execution details	56
Step Execution Progress	56
35. Analytics	58
VI. Appendices	59
A. Migrating from Spring XD to Spring Cloud Data Flow	60
A.1. Terminology Changes	60
A.2. Modules to Applications	60
Custom Applications	60
Application Registration	60
Application Properties	61
A.3. Message Bus to Binders	61
Message Bus	61
Binders	61
Named Channels	62
Directed Graphs	62
A.4. Batch to Tasks	62
A.5. Shell/DSL Commands	63
A.6. REST-API	63
A.7. UI / Flo	63
A.8. Architecture Components	64
ZooKeeper	64
RDBMS	64
Redis	64
Cluster Topology	64
A.9. Central Configuration	64
A.10. Distribution	64

A.11. Hadoop Distribution Compatibility	65
A.12. YARN Deployment	65
A.13. Use Case Comparison	65
Use Case #1	65
Use Case #2	66
Use Case #3	66
B. Test Cluster	68
B.1. Clone and Download files	68
B.2. Start Vagrant VMs for the Mesos cluster	68
B.3. Install Chronos as a Service running in Marathon	68
C. Building	70
C.1. Documentation	70
C.2. Working with the code	70
Importing into eclipse with m2eclipse	70
Importing into eclipse without m2eclipse	71
D. Contributing	72
D.1. Sign the Contributor License Agreement	72
D.2. Code Conventions and Housekeeping	72

Part I. Introduction

1. Introducing Spring Cloud Data Flow for Mesos and Marathon

This project provides support for orchestrating long-running (*streaming*) and short-lived (*task/batch*) data microservices to Marathon on Mesos.

2. Spring Cloud Data Flow

Spring Cloud Data Flow is a cloud-native orchestration service for composable data microservices on modern runtimes. With Spring Cloud Data Flow, developers can create and orchestrate data pipelines for common use cases such as data ingest, real-time analytics, and data import/export.

The Spring Cloud Data Flow architecture consists of a server that deploys [Streams](#) and [Tasks](#). Streams are defined using a [DSL](#) or visually through the browser based designer UI. Streams are based on the [Spring Cloud Stream](#) programming model while Tasks are based on the [Spring Cloud Task](#) programming model. The sections below describe more information about creating your own custom Streams and Tasks

For more details about the core architecture components and the supported features, please review Spring Cloud Data Flow's [core reference guide](#). There're several [samples](#) available for reference.

3. Spring Cloud Stream

Spring Cloud Stream is a framework for building message-driven microservice applications. Spring Cloud Stream builds upon Spring Boot to create standalone, production-grade Spring applications, and uses Spring Integration to provide connectivity to message brokers. It provides opinionated configuration of middleware from several vendors, introducing the concepts of persistent publish-subscribe semantics, consumer groups, and partitions.

For more details about the core framework components and the supported features, please review Spring Cloud Stream's [reference guide](#).

There's a rich ecosystem of Spring Cloud Stream [Application-Starters](#) that can be used either as standalone data microservice applications or in Spring Cloud Data Flow. For convenience, we have generated RabbitMQ and Apache Kafka variants of these application-starters that are available for use from [Maven Repo](#) and [Docker Hub](#) as maven artifacts and docker images, respectively.

Do you have a requirement to develop custom applications? No problem. Refer to this guide to create [custom stream applications](#). There're several [samples](#) available for reference.

4. Spring Cloud Task

Spring Cloud Task makes it easy to create short-lived microservices. We provide capabilities that allow short-lived JVM processes to be executed on demand in a production environment.

For more details about the core framework components and the supported features, please review Spring Cloud Task's [reference guide](#).

There's a rich ecosystem of Spring Cloud Task [Application-Starters](#) that can be used either as standalone data microservice applications or in Spring Cloud Data Flow. For convenience, the generated application-starters are available for use from [Maven Repo](#). There are several [samples](#) available for reference.

Part II. Architecture

5. Introduction

Spring Cloud Data Flow simplifies the development and deployment of applications focused on data processing use-cases. The major concepts of the architecture are Applications, the Data Flow Server, and the target runtime.

Applications come in two flavors

- Long lived Stream applications where an unbounded amount of data is consumed or produced via messaging middleware.
- Short lived Task applications that process a finite set of data and then terminate.

Depending on the runtime, applications can be packaged in two ways

- Spring Boot uber-jar that is hosted in a maven repository, file, http or any other Spring resource implementation.
- Docker

The runtime is the place where applications execute. The target runtimes for applications are platforms that you may already be using for other application deployments.

The supported runtimes are

- Cloud Foundry
- Apache YARN
- Kubernetes
- Apache Mesos
- Local Server for development

There is a deployer Service Provider Interface (SPI) that enables you to extend Data Flow to deploy onto other runtimes, for example to support Hashicorp's Nomad or Docker Swarm. Contributions are welcome!

The component that is responsible for deploying applications to a runtime is the Data Flow Server. There is a Data Flow Server executable jar provided for each of the target runtimes. The Data Flow server is responsible for interpreting

- A stream DSL that describes the logical flow of data through multiple applications.
- A deployment manifest that describes the mapping of applications onto the runtime. For example, to set the initial number of instances, memory requirements, and data partitioning.

As an example, the DSL to describe the flow of data from an http source to an Apache Cassandra sink would be written as "http | cassandra". These names in the DSL are registered with the Data Flow Server and map onto application artifacts that can be hosted in Maven or Docker repositories. Many source, processor, and sink applications for common use-cases (e.g. jdbc, hdfs, http, router) are provided by the Spring Cloud Data Flow team. The pipe symbol represents the communication between the two applications via messaging middleware. The two messaging middleware brokers that are supported are

- Apache Kafka
- RabbitMQ

In the case of Kafka, when deploying the stream, the Data Flow server is responsible to create the topics that correspond to each pipe symbol and configure each application to produce or consume from the topics so the desired flow of data is achieved.

The interaction of the main components is shown below

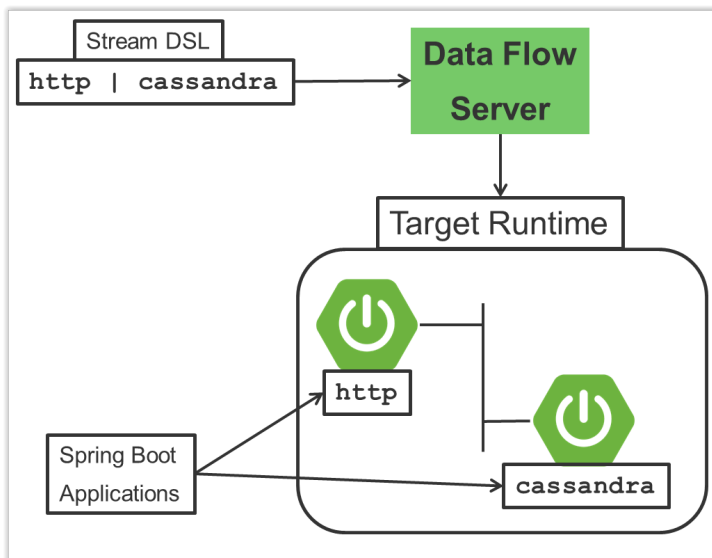


Figure 5.1. The Spring Cloud Data High Level Architecture

In this diagram a DSL description of a stream is POSTed to the Data Flow Server. Based on the mapping of DSL application names to Maven and Docker artifacts, the http source and cassandra sink application are deployed on the target runtime.

6. Microservice Architectural Style

The Data Flow Server deploys applications onto the target runtime that conform to the microservice architectural style. For example, a stream represents a high level application that consists of multiple small microservice applications each running in their own process. Each microservice application can be scaled up or down independent of the other and each has their own versioning lifecycle.

Both Streaming and Task based microservice applications build upon Spring Boot as the foundational library. This gives all microservice applications functionality such as health checks, security, configurable logging, monitoring and management functionality, as well as executable JAR packaging.

It is important to emphasise that these microservice applications are ‘just apps’ that you can run by yourself using ‘java -jar’ and passing in appropriate configuration properties. We provide many common microservice applications for common operations so you don’t have to start from scratch when addressing common use-cases which build upon the rich ecosystem of Spring Projects, e.g Spring Integration, Spring Data, Spring Hadoop and Spring Batch. Creating your own microservice application is similar to creating other Spring Boot applications, you can start using the Spring Initializr web site or the UI to create the basic scaffolding of either a Stream or Task based microservice.

In addition to passing in the appropriate configuration to the applications, the Data Flow server is responsible for preparing the target platform’s infrastructure so that the application can be deployed. For example, in Cloud Foundry it would be binding specified services to the applications and executing the ‘cf push’ command for each application. For Kubernetes it would be creating the replication controller, service, and load balancer.

The Data Flow Server helps simplify the deployment of multiple applications onto a target runtime, but one could also opt to deploy each of the microservice applications manually and not use Data Flow at all. This approach might be more appropriate to start out with for small scale deployments, gradually adopting the convenience and consistency of Data Flow as you develop more applications. Manual deployment of Stream and Task based microservices is also a useful educational exercise that will help you better understand some of the automatic applications configuration and platform targeting steps that the Data Flow Server provides.

6.1 Comparison to other Platform architectures

Spring Cloud Data Flow’s architectural style is different than other Stream and Batch processing platforms. For example in Apache Spark, Apache Flink, and Google Cloud Dataflow applications run on a dedicated compute engine cluster. The nature of the compute engine gives these platforms a richer environment for performing complex calculations on the data as compared to Spring Cloud Data Flow, but it introduces complexity of another execution environment that is often not needed when creating data centric applications. That doesn’t mean you cannot do real time data computations when using Spring Cloud Data Flow. Refer to the analytics section which describes the integration of Redis to handle common counting based use-cases as well as the RxJava integration for functional API driven analytics use-cases, such as time-sliding-window and moving-average among others.

Similarly, Apache Storm, Hortonworks DataFlow and Spring Cloud Data Flow’s predecessor, Spring XD, use a dedicated application execution cluster, unique to each product, that determines where your code should execute on the cluster and perform health checks to ensure that long lived applications are restarted if they fail. Often, framework specific interfaces are required to be used in order to correctly “plug in” to the cluster’s execution framework.

As we discovered during the evolution of Spring XD, the rise of multiple container frameworks in 2015 made creating our own runtime a duplication of efforts. There is no reason to build your own resource management mechanics, when there's multiple runtime platforms that offer this functionality already. Taking these considerations into account is what made us shift to the current architecture where we delegate the execution to popular runtimes, runtimes that you may already be using for other purposes. This is an advantage in that it reduces the cognitive distance for creating and managing data centric applications as many of the same skills used for deploying other end-user/web applications are applicable.

7. Streaming Applications

While Spring Boot provides the foundation for creating DevOps friendly microservice applications, other libraries in the Spring ecosystem help create Stream based microservice applications. The most important of these is Spring Cloud Stream.

The essence of the Spring Cloud Stream programming model is to provide an easy way to describe multiple inputs and outputs of an application that communicate over messaging middleware. These input and outputs map onto Kafka topics or Rabbit exchanges and queues. Common application configuration for a Source that generates data, a Process that consumes and produces data and a Sink that consumes data is provided as part of the library.

7.1 Imperative Programming Model

Spring Cloud Stream is most closely integrated with Spring Integration's imperative "event at a time" programming model. This means you write code that handles a single event callback. For example,

```
@EnableBinding(Sink.class)
public class LoggingSink {

    @StreamListener(Sink.INPUT)
    public void log(String message) {
        System.out.println(message);
    }
}
```

In this case the String payload of a message coming on the input channel, is handed to the log method. The `@EnableBinding` annotation is what is used to tie together the input channel to the external middleware.

7.2 Functional Programming Model

However, Spring Cloud Stream can support other programming styles. There is initial support for functional style programming via [RxJava Observable APIs](#) and upcoming versions will support callback methods with Project Reactor's Flux API and Apache Kafka's KStream API.

8. Streams

8.1 Topologies

The Stream DSL describes linear sequences of data flowing through the system. For example, in the stream definition `http | transformer | cassandra`, each pipe symbol connects the application on the left to the one on the right. Named channels can be used for routing and to fan out data to multiple messaging destinations.

Taps can be used to ‘listen in’ to the data that is flowing across any of the pipe symbols. Taps can be used as sources for new streams with an independent life cycle.

8.2 Concurrency

For an application that will consume events, Spring Cloud Stream exposes a concurrency setting that controls the size of a thread pool used for dispatching incoming messages. See the [Consumer properties](#) documentation for more information.

8.3 Partitioning

A common pattern in stream processing is to partition the data as it moves from one application to the next. Partitioning is a critical concept in stateful processing, for either performance or consistency reasons, to ensure that all related data is processed together. For example, in a time-windowed average calculation example, it is important that all measurements from any given sensor are processed by the same application instance. Alternatively, you may want to cache some data related to the incoming events so that it can be enriched without making a remote procedure call to retrieve the related data.

Spring Cloud Data Flow supports partitioning by configuring Spring Cloud Stream’s output and input bindings. Spring Cloud Stream provides a common abstraction for implementing partitioned processing use cases in a uniform fashion across different types of middleware. Partitioning can thus be used whether the broker itself is naturally partitioned (e.g., Kafka topics) or not (e.g., RabbitMQ). The following image shows how data could be partitioned into two buckets, such that each instance of the average processor application consumes a unique set of data.

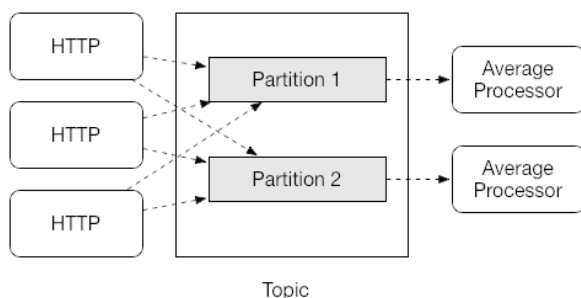


Figure 8.1. Spring Cloud Stream Partitioning

To use a simple partitioning strategy in Spring Cloud Data Flow, you only need set the instance count for each application in the stream and a `partitionKeyExpression` producer property when deploying the stream. The `partitionKeyExpression` identifies what part of the message will be used as the key to partition data in the underlying middleware. An `ingest` stream can be defined as `http | averageprocessor | cassandra` (Note that the Cassandra sink isn’t shown in the diagram above).

Suppose the payload being sent to the http source was in JSON format and had a field called `sensorId`. Deploying the stream with the shell command `stream deploy ingest --propertiesFile ingestStream.properties` where the contents of the file `ingestStream.properties` are

```
app.http.count=3
app.averageprocessor.count=2
app.http.producer.partitionKeyExpression=payload.sensorId
```

will deploy the stream such that all the input and output destinations are configured for data to flow through the applications but also ensure that a unique set of data is always delivered to each `averageprocessor` instance. In this case the default algorithm is to evaluate `payload.sensorId % partitionCount` where the `partitionCount` is the application count in the case of RabbitMQ and the partition count of the topic in the case of Kafka.

Please refer to [the section called “Passing stream partition properties during stream deployment”](#) for additional strategies to partition streams during deployment and how they map onto the underlying [Spring Cloud Stream Partitioning properties](#).

Also note, that you can’t currently scale partitioned streams. Read the section [Section 12.3, “Scaling at runtime”](#) for more information.

8.4 Message Delivery Guarantees

For consumer applications, there is a retry policy for exceptions generated during message handling. The default is to retry the callback method invocation 3 times and wait one second for the first retry. A backoff multiplier of 2 is used for the second and third attempts. All of these retry properties are configurable.

If there is still an exception on the last retry attempt, and dead letter queues are enabled, the message and exception message are published to the dead letter queue. The dead letter queue is a destination and its nature depends on the messaging middleware (e.g in the case of Kafka it is a dedicated topic). If dead letter functionality is not enabled, the message and exception is sent to the error channel, which by default logs the message and exception.

Additional messaging delivery guarantees are those provided by the underlying messaging middleware that is chosen for the application for both producing and consuming applications. Refer to the Kafka [Consumer](#) and [Producer](#) and Rabbit [Consumer](#) and [Producer](#) documentation for more details. You will find there to be extensive declarative support for all the native QOS options.

9. Analytics

Spring Cloud Data Flow is aware of certain Sink applications that will write counter data to Redis and provides an REST endpoint to read counter data. The types of counters supported are

- [Counter](#) - Counts the number of messages it receives, optionally storing counts in a separate store such as redis.
- [Field Value Counter](#) - Counts occurrences of unique values for a named field in a message payload
- [Aggregate Counter](#) - Stores total counts but also retains the total count values for each minute, hour day and month.

It is important to note that the timestamp that is used in the aggregate counter can come from a field in the message itself so that out of order messages are properly accounted.

10. Task Applications

The Spring Cloud Task programming model provides:

- Persistence of the Task's lifecycle events and exit code status.
- Lifecycle hooks to execute code before or after a task execution.
- Emit task events to a stream (as a source) during the task lifecycle.
- Integration with Spring Batch Jobs.

11. Data Flow Server

11.1 Endpoints

The Data Flow Server uses an embedded servlet container and exposes REST endpoints for creating, deploying, undeploying, and destroying streams and tasks, querying runtime state, analytics, and the like. The Data Flow Server is implemented using Spring's MVC framework and the [Spring HATEOAS](#) library to create REST representations that follow the HATEOAS principle.

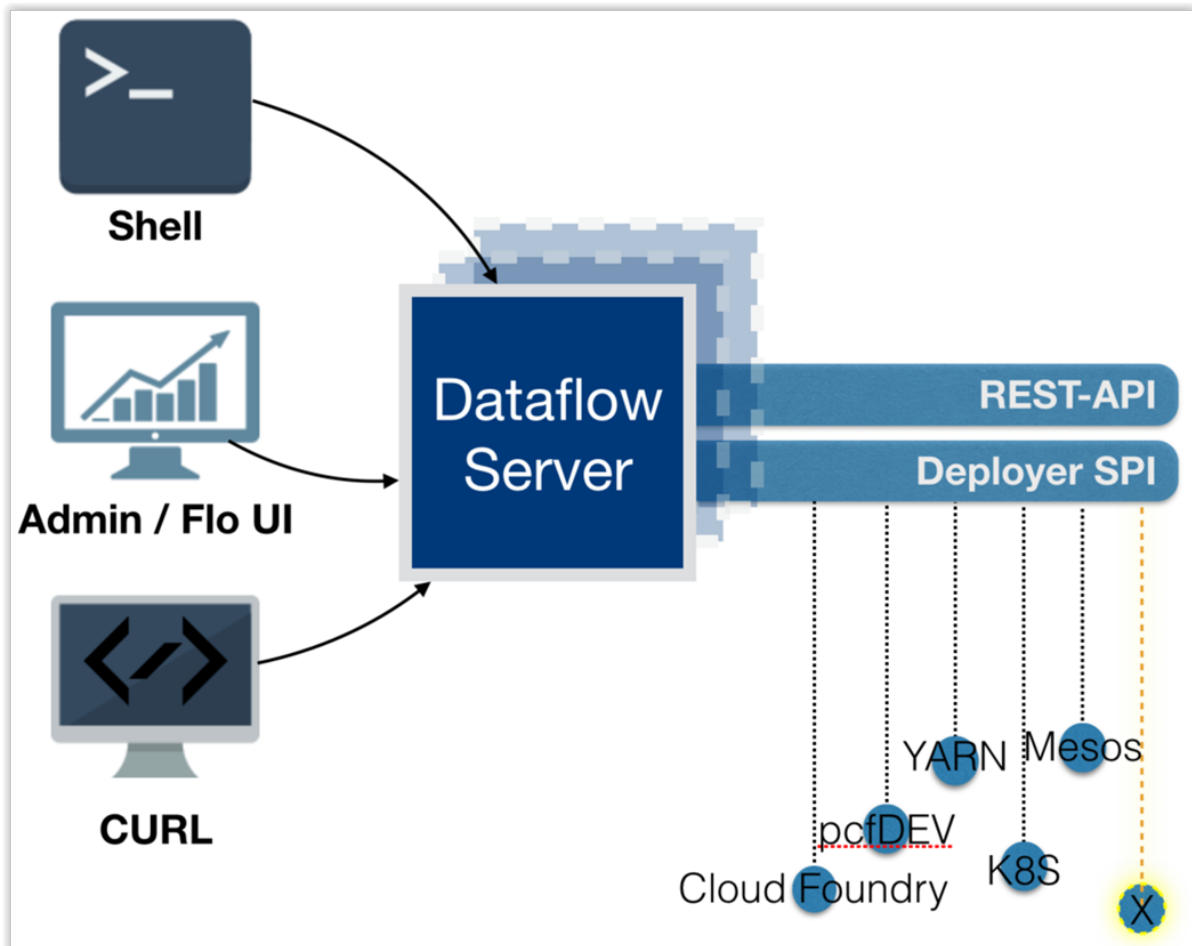


Figure 11.1. The Spring Cloud Data Flow Server

11.2 Customization

Each Data Flow Server executable jar targets a single runtime by delegating to the implementation of the deployer Service Provider Interface found on the classpath.

We provide a Data Flow Server executable jar that targets a single runtime. The Data Flow server delegates to the implementation of the deployer Service Provider Interface found on the classpath. In the current version, there are no endpoints specific to a target runtime, but may be available in future releases as a convenience to access runtime specific features

While we provide a server executable for each of the target runtimes you can also create your own customized server application using Spring Initializr. This let's you add or remove functionality relative to the executable jar we provide. For example, adding additional security implementations, custom

endpoints, or removing Task or Analytics REST endpoints. You can also enable or disable some features through the use of feature toggles.

11.3 Security

The Data Flow Server executable jars support basic http and OAuth 2.0 authentication to access it endpoints. Refer to the security section for more information.

Authorization via groups is planned for a future release.

12. Runtime

12.1 Fault Tolerance

The target runtimes supported by Data Flow all have the ability to restart a long lived application should it fail. Spring Cloud Data Flow sets up whatever health probe is required by the runtime environment when deploying the application.

The collective state of all applications that comprise the stream is used to determine the state of the stream. If an application fails, the state of the stream will change from 'deployed' to 'partial'.

12.2 Resource Management

Each target runtime lets you control the amount of memory, disk and CPU that is allocated to each application. These are passed as properties in the deployment manifest using key names that are unique to each runtime. Refer to the each platforms server documentation for more information.

12.3 Scaling at runtime

When deploying a stream, you can set the instance count for each individual application that comprises the stream. Once the stream is deployed, each target runtime lets you control the target number of instances for each individual application. Using the APIs, UIs, or command line tools for each runtime, you can scale up or down the number of instances as required. Future work will provide a portable command in the Data Flow Server to perform this operation.

Currently, this is not supported with the Kafka binder (based on the 0.8 simple consumer at the time of the release), as well as partitioned streams, for which the suggested workaround is redeploying the stream with an updated number of instances. Both cases require a static consumer set up based on information about the total instance count and current instance index, a limitation intended to be addressed in future releases. For example, Kafka 0.9 and higher provides good infrastructure for scaling applications dynamically and will be available as an alternative to the current Kafka 0.8 based binder in the near future. One specific concern regarding scaling partitioned streams is the handling of local state, which is typically reshuffled as the number of instances is changed. This is also intended to be addressed in the future versions, by providing first class support for local state management.

12.4 Application Versioning

Application versioning, that is upgrading or downgrading an application from one version to another, is not directly supported by Spring Cloud Data Flow. You must rely on specific target runtime features to perform these operational tasks.

The roadmap for Spring Cloud Data Flow will deploy applications that are compatible with Spinnaker to manage the complete application lifecycle. This also includes automated canary analysis backed by application metrics. Portable commands in the Data Flow server to trigger pipelines in Spinnaker are also planned.

Part III. Getting Started

13. Deploying Streams and Tasks on Mesos and Marathon/Chronos

In this getting started the Data Flow Server is running as a Docker container in Marathon and so are all the dependent services like a relational database for stream and task repositories, a message bus for stream apps and the key value store for analytics.

1. Deploy a Mesos and Marathon cluster.

The [Mesosphere getting started guide](#) provides a number of options for you to deploy a cluster. There is also a number of options listed on Mesosphere's [Install DC/OS](#) page. In [Appendix B, Test Cluster](#) we describe how we configured a local test cluster using the DC/OS Vagrant project.

The rest of this getting started guide assumes that you have a working Mesos and Marathon cluster and know the Marathon endpoint URL.

We are using the Marathon endpoint URL of <http://m1.dcos/service/marathon> for this document.

2. Create a MySQL service on the Mesos cluster.

The `mysql` service will be used for storing stream and task definitions in the stream and task repositories. There is a sample [application JSON file for MySQL](#) in the `spring-cloud-dataflow-server-mesos` repository that you can use as a starting point. The service discovery mechanism is currently disabled so you need to look up the host and port to use for the connection. Depending on how large your cluster is, you may want to tweak the CPU and/or memory values.

Using the above JSON file and an Mesos and Marathon cluster installed you can deploy a Rabbit MQ application instance by issuing the following command

```
curl -X POST http://m1.dcos/service/marathon/v2/apps -d @mysql.json -H "Content-type: application/json"
```



Note

Note the `@` symbol to reference a file input for the `curl` command.

3. Create a Rabbit MQ service on the Mesos cluster.

The `rabbitmq` service will be used for messaging between applications in the stream. There is a sample [application JSON file for Rabbit MQ](#) in the `spring-cloud-dataflow-server-mesos` repository that you can use as a starting point. The service discovery mechanism is currently disabled so you need to look up the host and port to use for the connection. Depending on how large your cluster is, you may want to tweak the CPU and/or memory values.

Using the above JSON file and an Mesos and Marathon cluster installed you can deploy a Rabbit MQ service instance by issuing the following command

```
curl -X POST http://m1.dcos/service/marathon/v2/apps -d @rabbitmq.json -H "Content-type: application/json"
```

4. Create a Redis service on the Mesos cluster.

The `redis` service will be used for counters as part of the analytics. There is a sample [application JSON file for Redis](#) in the `spring-cloud-dataflow-server-mesos` repository that you can use

as a starting point. The service discovery mechanism is currently disabled so you need to look up the host and port to use for the connection. Depending on how large your cluster is, you may want to tweak the CPU and/or memory values.

Using the above JSON file and an Mesos and Marathon cluster installed you can deploy a Redis service instance by issuing the following command

```
curl -X POST http://ml.dcos/service/marathon/v2/apps -d @redis.json -H "Content-type: application/json"
```

Using the Marathon and Mesos UIs you can verify that `mysql`, `rabbitmq` and `redis` services are running on the cluster.

5. Install Chronos on the Mesos cluster.

The [Chronos](#) service will be used for running task. If you haven't already installed Chronos, this is the time to do it. You can install Chronos using the DC/OS UI (under the Universe section) or from the `dcos` command line:

```
dcos package install chronos
```

6. Download the Marathon application JSON for the Spring Cloud Data Flow Server.

Use the following command to download the Marathon application JSON file used to deploy Spring Cloud Data Flow Server for Mesos.

```
$ wget https://raw.githubusercontent.com/spring-cloud/spring-cloud-dataflow-server-mesos/master/src/etc/marathon/scdf-server.json
```

We will need to modify the Docker image tag, API endpoints and host and port settings based on the current deployment environment. We hope to eliminate most of this in future releases and instead rely on service discovery mechanisms. For now we do have to make the modifications manually. The downloaded file should look like this:

```
{
  "id": "/spring-cloud-data-flow",
  "cpus": 0.5,
  "mem": 512.0,
  "instances": 1,
  "container": {
    "type": "DOCKER",
    "docker": {
      "image": "springcloud/spring-cloud-dataflow-server-mesos",
      "network": "BRIDGE",
      "portMappings": [
        { "containerPort": 9393 }
      ]
    }
  },
  "env": {
    "MESOS_MARATHON_URI": "http://ml.dcos/service/marathon",
    "MESOS_CHRONOS_URI": "http://ml.dcos/service/chronos",
    "JDBC_URL": "jdbc:mysql://192.168.65.111:5769/test",
    "JDBC_DRIVER": "org.mariadb.jdbc.Driver",
    "JDBC_USERNAME": "spring",
    "JDBC_PASSWORD": "secret",
    "RABBITMQ_HOST": "192.168.65.121",
    "RABBITMQ_PORT": "5261",
    "REDIS_HOST": "192.168.65.111",
    "REDIS_PORT": "19902",
    "SPRING_APPLICATION_JSON": "{\"spring.cloud.deployer.mesos.marathon.apiEndpoint\": \"${MESOS_MARATHON_URI}\", \"spring.cloud.deployer.mesos.chronos.apiEndpoint\"
```

```
\": \"${MESOS_CHRONOS_URI}\", \"spring.datasource.url\": \"${JDBC_URL}\",
\"spring.datasource.driverClassName\": \"${JDBC_DRIVER}\", \"spring.datasource.username
\": \"${JDBC_USERNAME}\", \"spring.datasource.password\": \"${JDBC_PASSWORD}\",
\"spring.datasource.testOnBorrow\": true, \"spring.datasource.validationQuery\": \"SELECT
1\", \"spring.redis.host\": \"${REDIS_HOST}\", \"spring.redis.port\": \"${REDIS_PORT}\",
\"spring.cloud.deployer.mesos.marathon.environmentVariables\": \"SPRING_RABBITMQ_HOST=
${RABBITMQ_HOST}, SPRING_RABBITMQ_PORT=${RABBITMQ_PORT}\"}
},
\"healthChecks\": [
{
\"path\": \"/management/health\",
\"portIndex\": 0,
\"protocol\": \"HTTP\",
\"ignoreHttp1xx\": false,
\"gracePeriodSeconds\": 120,
\"intervalSeconds\": 60,
\"timeoutSeconds\": 20,
\"maxConsecutiveFailures\": 0
}
]
}
```

First we need to modify the Docker image to use the tag `latest`. It should be:

```
"image": "springcloud/spring-cloud-dataflow-server-mesos:latest",
```

In the `env` section there are several environment variables that we need to adjust. We need to provide the following properties for accessing Marathon and Chronos APIs:

```
"MESOS_MARATHON_URI": "http://ml.dcos/service/marathon",
"MESOS_CHRONOS_URI": "http://ml.dcos/service/chronos",
```

Here we did set them to the defaults for a local Vagrant DC/OS installation.

We also need to provide the database configuration properties. Look up the database host and port from the Marathon UI. For the `mysql` service that we just installed, they were `192.168.65.111:5769`.

```
"JDBC_URL": "jdbc:mysql://192.168.65.111:5769/test",
"JDBC_DRIVER": "org.mariadb.jdbc.Driver",
"JDBC_USERNAME": "spring",
"JDBC_PASSWORD": "secret",
```

Next, we need to provide the message bus configuration properties. Look up the host and port for the `rabbitmq` service. In our case they were `192.168.65.121:5261`.

```
"RABBITMQ_HOST": "192.168.65.121",
"RABBITMQ_PORT": "5261",
```

Finally we need to do the same for the key value store properties. Look up the host and port for the `redis` service. In our case they were `192.168.65.111:19902`.

```
"REDIS_HOST": "192.168.65.111",
"REDIS_PORT": "19902",
```

You can add properties to the `SPRING_APPLICATION_JSON` property as well. You might want to set default values for memory and cpu resource request. For example `\\"spring.cloud.deployer.mesos.marathon.memory\\"=\\\"768\\"` will by default allocate

additional memory for the application vs. the default value of 512. You can see all the available options in the [MarathonAppDeployerProperties.java](#) file.



Note

DC/OS in secured mode requires an Authorization header with a token when accessing the Marathon and Chronos REST end-points. To accommodate this you need to provide this token when deploying the Spring Cloud Data Flow server to a DC/OS secured cluster. See below for instructions.

If you are using a secured DC/OS cluster then you will need to add the authorization token to the above configuration. First, log in using `dcos auth login` command and enter the authentication token you get after authenticating with DC/OS web interface with the provided link. Next, run the `dcos config show core.dcos_acs_token` command to display the authorization token we need for our configuration. Copy and paste this token in the following environment variable that you add to the above application JSON:

```
"DCOS_TOKEN": "<paste the token here>",
```

We also need to add the `spring.cloud.deployer.mesos.dcos.authorizationToken` property to the `SPRING_APPLICATION_JSON` entry. Insert the following as another property entry:

```
,{"spring.cloud.deployer.mesos.dcos.authorizationToken\":\"${DCOS_TOKEN}\"}
```

- Now, deploy the Spring Cloud Data Flow Server for Mesos and Marathon/Chronos using the above modified application JSON.

```
curl -X POST http://ml.dcos/service/marathon/v2/apps -d @scdf-server.json -H "Content-type: application/json"
```

Verify that the `spring-cloud-data-flow` application is running before proceeding.

- Download and run the Spring Cloud Data Flow shell.

```
$ wget http://repo.spring.io/release/org/springframework/cloud/spring-cloud-dataflow-shell/1.0.1.RELEASE/spring-cloud-dataflow-shell-1.0.1.RELEASE.jar
$ java -jar spring-cloud-dataflow-shell-1.0.1.RELEASE.jar
```

Lookup the host and port for the `spring-cloud-data-flow` application in the Marathon UI. Use those values to configure the server URI for the shell:

```
dataflow:>dataflow config server --uri http://192.168.65.111:20043
```

- By default, the application registry will be empty. If you would like to register all out-of-the-box stream applications built with the RabbitMQ binder in bulk, you can with the following command. For more details, review how to [register applications](#).

```
dataflow:>app import --uri http://bit.ly/stream-applications-rabbit-docker
```

- Deploy a simple stream in the shell



Note

If you need to specify any of the app specific configuration properties then you must use "long-form" of them including the app specific prefix like `--jdbc.tableName=TEST_DATA`.

This is due to the server not being able to access the metadata for the Docker based starter apps. You will also not see the configuration properties listed when using the `app info` command or in the Dashboard GUI.

```
dataflow:>stream create --name ticktock --definition "time | log" --deploy
```

In the Mesos UI you can then look at the logs for the log sink. Look for a Mesos task with the name `log-0.log.ticktock`.

```
2016-04-26 18:13:03.001 INFO 1 --- [          main] s.b.c.e.t.TomcatEmbeddedServletContainer :
Tomcat started on port(s): 8080 (http)
2016-04-26 18:13:03.004 INFO 1 --- [          main] o.s.c.s.a.l.s.r.LogSinkRabbitApplication :
Started LogSinkRabbitApplication in 7.766 seconds (JVM running for 8.24)
2016-04-26 18:13:54.443 INFO 1 --- [nio-8080-exec-1] o.a.c.c.C.[Tomcat].[localhost].[/] :
Initializing Spring FrameworkServlet 'dispatcherServlet'
2016-04-26 18:13:54.445 INFO 1 --- [nio-8080-exec-1] o.s.web.servlet.DispatcherServlet :
FrameworkServlet 'dispatcherServlet': initialization started
2016-04-26 18:13:54.459 INFO 1 --- [nio-8080-exec-1] o.s.web.servlet.DispatcherServlet :
FrameworkServlet 'dispatcherServlet': initialization completed in 14 ms
2016-04-26 18:14:09.088 INFO 1 --- [time.ticktock-1] log.sink :
04/26/16 18:14:09
2016-04-26 18:14:10.077 INFO 1 --- [time.ticktock-1] log.sink :
04/26/16 18:14:10
2016-04-26 18:14:11.080 INFO 1 --- [time.ticktock-1] log.sink :
04/26/16 18:14:11
2016-04-26 18:14:12.083 INFO 1 --- [time.ticktock-1] log.sink :
04/26/16 18:14:12
2016-04-26 18:14:13.090 INFO 1 --- [time.ticktock-1] log.sink :
04/26/16 18:14:13
2016-04-26 18:14:14.091 INFO 1 --- [time.ticktock-1] log.sink :
04/26/16 18:14:14
2016-04-26 18:14:15.093 INFO 1 --- [time.ticktock-1] log.sink :
04/26/16 18:14:15
2016-04-26 18:14:16.095 INFO 1 --- [time.ticktock-1] log.sink :
04/26/16 18:14:16
```

11 Destroy the stream

```
dataflow:>stream destroy --name ticktock
```

12 Register a task application using the shell

```
dataflow:>app register --name timestamp --type task --uri docker:springcloudtask/timestamp-
task:latest
```

13 Create and launch the task using the shell

```
dataflow:>task create testtask --definition "timestamp"
dataflow:>task launch testtask
```

In the Mesos UI you can then look at the logs for the `testtask` task. Look for a Mesos task with the name `ChronosTask:testtask`.

```
Starting task ct:1472062219364:0:testtask:
```

```
.
/\ /  _ _ _ _ _ ( _ ) _ _ _ _ \ \ \ \
( ( ) _ _ | ' _ | _ | ' _ \ _ | \ \ \ \
\ \ / _ _ | ( _ ) | | | | | ( _ ) | ) ) )
' | _ _ | . _ | | _ | | _ \ , / / / /
=====|_|=====|_|/_/_/_/_/
:: Spring Boot ::                (v1.3.5.RELEASE)
```

```

2016-08-24 18:10:45.957 INFO 1 --- [          main] o.s.c.t.a.t.TimestampTaskApplication :
Starting TimestampTaskApplication v1.0.2.BUILD-SNAPSHOT on a2.dcos with PID 1 (/maven/timestamp-
task.jar started by root in /)
2016-08-24 18:10:45.960 INFO 1 --- [          main] o.s.c.t.a.t.TimestampTaskApplication : No
active profile set, falling back to default profiles: default
2016-08-24 18:10:46.003 INFO 1 --- [          main] s.c.a.AnnotationConfigApplicationContext :
Refreshing org.springframework.context.annotation.AnnotationConfigApplicationContext@788c6159:
startup date [Wed Aug 24 18:10:46 GMT 2016]; root of context hierarchy
2016-08-24 18:10:47.051 INFO 1 --- [          main] o.s.jdbc.datasource.init.ScriptUtils :
Executing SQL script from class path resource [org/springframework/cloud/task/schema-mysql.sql]
2016-08-24 18:10:47.062 INFO 1 --- [          main] o.s.jdbc.datasource.init.ScriptUtils :
Executed SQL script from class path resource [org/springframework/cloud/task/schema-mysql.sql] in 11
ms.
2016-08-24 18:10:47.207 INFO 1 --- [          main] o.s.j.e.a.AnnotationMBeanExporter :
Registering beans for JMX exposure on startup
2016-08-24 18:10:47.211 INFO 1 --- [          main] o.s.c.support.DefaultLifecycleProcessor :
Starting beans in phase 0
2016-08-24 18:10:47.238 INFO 1 --- [          main] TimestampTaskConfiguration$TimestampTask :
2016-08-24 18:10:47.238
2016-08-24 18:10:47.249 INFO 1 --- [          main] s.c.a.AnnotationConfigApplicationContext :
Closing org.springframework.context.annotation.AnnotationConfigApplicationContext@788c6159: startup
date [Wed Aug 24 18:10:46 GMT 2016]; root of context hierarchy
2016-08-24 18:10:47.250 INFO 1 --- [          main] o.s.c.support.DefaultLifecycleProcessor :
Stopping beans in phase 0
2016-08-24 18:10:47.252 INFO 1 --- [          main] o.s.j.e.a.AnnotationMBeanExporter :
Unregistering JMX-exposed beans on shutdown
2016-08-24 18:10:47.261 INFO 1 --- [          main] o.s.c.t.a.t.TimestampTaskApplication :
Started TimestampTaskApplication in 1.62 seconds (JVM running for 2.018)```

```

14 Destroy the task

```
dataflow:>task destroy --name testtask
```

Part IV. Streams

In this section you will learn all about Streams and how to use them with Spring Cloud Data Flow.

14. Introduction

In Spring Cloud Data Flow, a basic stream defines the ingestion of event driven data from a *source* to a *sink* that passes through any number of *processors*. Streams are composed of spring-cloud-stream applications and the deployment of stream definitions is done via the Data Flow Server (REST API). The [Getting Started](#) section shows you how to start these servers and how to start and use the Spring Cloud Data Flow shell.

A high level DSL is used to create stream definitions. The DSL to define a stream that has an http source and a file sink (with no processors) is shown below

```
http | file
```

The DSL mimics a UNIX pipes and filters syntax. Default values for ports and filenames are used in this example but can be overridden using `--` options, such as

```
http --server.port=8091 | file --directory=/tmp/httpdata/
```

To create these stream definitions you use the shell or make an HTTP POST request to the Spring Cloud Data Flow Server. More details can be found in the sections below.

15. Stream DSL

In the examples above, we connected a source to a sink using the pipe symbol `|`. You can also pass properties to the source and sink configurations. The property names will depend on the individual app implementations, but as an example, the `http` source app exposes a `server.port` setting which allows you to change the data ingestion port from the default value. To create the stream using port 8000, we would use

```
dataflow:> stream create --definition "http --server.port=8000 | log" --name myhttpstream
```

The shell provides tab completion for application properties and also the shell command `app info` provides some additional documentation.

16. Register a Stream App

Register a Stream App with the App Registry using the Spring Cloud Data Flow Shell `app register` command. You must provide a unique name, application type, and a URI that can be resolved to the app artifact. For the type, specify "source", "processor", or "sink". Here are a few examples:

```
dataflow:>app register --name mysource --type source --uri maven://com.example:mysource:0.0.1-SNAPSHOT

dataflow:>app register --name myprocessor --type processor --uri file:///Users/example/
myprocessor-1.2.3.jar

dataflow:>app register --name mysink --type sink --uri http://example.com/mysink-2.0.1.jar
```

When providing a URI with the maven scheme, the format should conform to the following:

```
maven://<groupId>:<artifactId>[:<extension>[:<classifier>]]:<version>
```

For example, if you would like to register the snapshot versions of the `http` and `log` applications built with the RabbitMQ binder, you could do the following:

```
dataflow:>app register --name http --type source --uri maven://
org.springframework.cloud.stream.app:http-source-rabbit:1.0.0.BUILD-SNAPSHOT
dataflow:>app register --name log --type sink --uri maven://org.springframework.cloud.stream.app:log-
sink-rabbit:1.0.0.BUILD-SNAPSHOT
```

If you would like to register multiple apps at one time, you can store them in a properties file where the keys are formatted as `<type>.<name>` and the values are the URIs.

For example, if you would like to register the snapshot versions of the `http` and `log` applications built with the RabbitMQ binder, you could have the following in a properties file [eg: *stream-apps.properties*]:

```
source.http=maven://org.springframework.cloud.stream.app:http-source-rabbit:1.0.0.BUILD-SNAPSHOT
sink.log=maven://org.springframework.cloud.stream.app:log-sink-rabbit:1.0.0.BUILD-SNAPSHOT
```

Then to import the apps in bulk, use the `app import` command and provide the location of the properties file via `--uri`:

```
dataflow:>app import --uri file:///<YOUR_FILE_LOCATION>/stream-apps.properties
```

For convenience, we have the static files with application-URIs (for both maven and docker) available for all the out-of-the-box Stream app-starters. You can point to this file and import all the application-URIs in bulk. Otherwise, as explained in previous paragraphs, you can register them individually or have your own custom property file with only the required application-URIs in it. It is recommended, however, to have a "focused" list of desired application-URIs in a custom property file.

List of available static property files:

- Maven based Stream Applications with RabbitMQ Binder: bit.ly/stream-applications-rabbit-maven
- Maven based Stream Applications with Kafka Binder: bit.ly/stream-applications-kafka-maven
- Docker based Stream Applications with RabbitMQ Binder: bit.ly/stream-applications-rabbit-docker
- Docker based Stream Applications with Kafka Binder: bit.ly/stream-applications-kafka-docker

For example, if you would like to register all out-of-the-box stream applications built with the RabbitMQ binder in bulk, you can with the following command.

```
dataflow:>app import --uri http://bit.ly/stream-applications-rabbit-maven
```

You can also pass the `--local` option (which is `TRUE` by default) to indicate whether the properties file location should be resolved within the shell process itself. If the location should be resolved from the Data Flow Server process, specify `--local false`.

When using either `app register` or `app import`, if a stream app is already registered with the provided name and type, it will not be overridden by default. If you would like to override the pre-existing stream app, then include the `--force` option.



Note

In some cases the Resource is resolved on the server side, whereas in others the URI will be passed to a runtime container instance where it is resolved. Consult the specific documentation of each Data Flow Server for more detail.

16.1 Whitelisting application properties

Stream applications are Spring Boot applications which are aware of many [common application properties](#), e.g. `server.port` but also families of properties such as those with the prefix `spring.jmx` and `logging`. When creating your own application it is desirable to whitelist properties so that the shell and the UI can display them first as primary properties when presenting options via TAB completion or in drop-down boxes.

To whitelist application properties create a file named `spring-configuration-metadata-whitelist.properties` in the `META-INF` resource directory. There are two property keys that can be used inside this file. The first key is named `configuration-properties.classes`. The value is a comma separated list of fully qualified `@ConfigurationProperty` class names. The second key is `configuration-properties.names` whose value is a comma separated list of property names. This can contain the full name of property, such as `server.port` or a partial name to whitelist a category of property names, e.g. `spring.jmx`.

The [Spring Cloud Stream application starters](#) are a good place to look for examples of usage. Here is a simple example of the file source's `spring-configuration-metadata-whitelist.properties` file

```
configuration.classes=org.springframework.cloud.stream.app.file.sink.FileSinkProperties
```

If for some reason we also wanted to add `file.prefix` to this file, it would look like

```
configuration.classes=org.springframework.cloud.stream.app.file.sink.FileSinkProperties
configuration-properties.names=server.port
```



Important

As of Spring Cloud Data Flow 1.0.0.RELEASE the whitelisting of application properties is only explicitly supported for Spring Boot 1.3.x based application. Milestone releases of the upcoming Spring Boot 1.4.0 release are not explicitly supported, yet.

The `spring-boot-maven-plugin` used in 1.4.x has a different approach in handling the nested archives inside the jar. As a result you will notice that the application properties are not listed using `app info` command at all. As a temporary workaround, you can override the managed version of your app's `spring-boot-maven-plugin` explicitly and revert to a version of the latest 1.3.x release:

For example, if your app's `pom.xml` specifies to use Spring Boot 1.4.0.M3:

```
<parent>
  <artifactId>spring-boot-starter-parent</artifactId>
  <groupId>org.springframework.boot</groupId>
  <version>1.4.0.M3</version>
  <relativePath></relativePath>
</parent>
```

Then you can override the managed version of the `spring-boot-maven-plugin` with:

```
<plugin>
  <groupId>org.springframework.boot</groupId>
  <artifactId>spring-boot-maven-plugin</artifactId>
  <version>1.3.5.RELEASE</version> ❶
</plugin>
```

❶ Overriding the managed version 1.4.0.M3.

Also, if you have your own dataflow server built using `@EnableDataflowServer` and using Spring Boot 1.4.x in that, you would need to explicitly override the `spring-boot-maven-plugin` with any of 1.3.x releases.

17. Creating a Stream

The Spring Cloud Data Flow Server exposes a full RESTful API for managing the lifecycle of stream definitions, but the easiest way to use it is via the Spring Cloud Data Flow shell. Start the shell as described in the [Getting Started](#) section.

New streams are created by posting stream definitions. The definitions are built from a simple DSL. For example, let's walk through what happens if we execute the following shell command:

```
dataflow:> stream create --definition "time | log" --name ticktock
```

This defines a stream named `ticktock` based off the DSL expression `time | log`. The DSL uses the "pipe" symbol `|`, to connect a source to a sink.

Then to deploy the stream execute the following shell command (or alternatively add the `--deploy` flag when creating the stream so that this step is not needed):

```
dataflow:> stream deploy --name ticktock
```

The Data Flow Server resolves `time` and `log` to maven coordinates and uses those to launch the `time` and `log` applications of the stream.

```
2016-06-01 09:41:21.728 INFO 79016 --- [nio-9393-exec-6] o.s.c.d.spi.local.LocalAppDeployer :
    deploying app ticktock.log instance 0
    Logs will be in /var/folders/wn/8jxm_tbdlvj28c8vj37n900m0000gn/T/spring-cloud-
dataflow-912434582726479179/ticktock-1464788481708/ticktock.log
2016-06-01 09:41:21.914 INFO 79016 --- [nio-9393-exec-6] o.s.c.d.spi.local.LocalAppDeployer :
    deploying app ticktock.time instance 0
    Logs will be in /var/folders/wn/8jxm_tbdlvj28c8vj37n900m0000gn/T/spring-cloud-
dataflow-912434582726479179/ticktock-1464788481910/ticktock.time
```

In this example, the `time` source simply sends the current time as a message each second, and the `log` sink outputs it using the logging framework. You can tail the `stdout` log (which has an "`_instance>`" suffix). The log files are located within the directory displayed in the Data Flow Server's log output, as shown above.

```
$ tail -f /var/folders/wn/8jxm_tbdlvj28c8vj37n900m0000gn/T/spring-cloud-dataflow-912434582726479179/
ticktock-1464788481708/ticktock.log/stdout_0.log
2016-06-01 09:45:11.250 INFO 79194 --- [ kafka-binder-] log.sink : 06/01/16 09:45:11
2016-06-01 09:45:12.250 INFO 79194 --- [ kafka-binder-] log.sink : 06/01/16 09:45:12
2016-06-01 09:45:13.251 INFO 79194 --- [ kafka-binder-] log.sink : 06/01/16 09:45:13
```

17.1 Application properties

Application properties are the properties associated with each application in the stream. When the application is deployed, the application properties are applied to the application via command line arguments or environment variables based on the underlying deployment implementation.

Passing application properties when creating a stream

The following stream

```
dataflow:> stream create --definition "time / log" --name ticktock
```

can have application properties defined at the time of stream creation.

The shell command `app info` displays the white-listed application properties for the application. For more info on the property white listing refer to [Section 16.1, “Whitelisting application properties”](#)

Below are the white listed properties for the app `time`:

```
dataflow:> app info source:time
#####
#      Option Name      #      Description      #      Default      #
#      Type              #
#####
#trigger.time-unit      #The TimeUnit to apply to delay#<none>
#java.util.concurrent.TimeUnit #
#      #                  #values.                  #
#
#trigger.fixed-delay    #Fixed delay for periodic    #1
#java.lang.Integer      #
#      #                  #triggers.                  #
#
#trigger.cron            #Cron expression value for the #<none>
#java.lang.String       #
#      #                  #Cron Trigger.            #
#
#trigger.initial-delay  #Initial delay for periodic  #0
#java.lang.Integer      #
#      #                  #triggers.                  #
#
#trigger.max-messages   #Maximum messages per poll, -1 #1
#java.lang.Long          #
#      #                  #means infinity.          #
#
#trigger.date-format    #Format for the date value.  #<none>
#java.lang.String       #
#####
```

Below are the white listed properties for the app `log`:

```
dataflow:> app info sink:log
#####
#      Option Name      #      Description      #      Default      #
#      Type              #
#####
#log.name               #The name of the logger to use.#<none>
#java.lang.String       #
#log.level              #The level at which to log    #<none>
#org.springframework.integration#
#      #                  #messages.                  #
#n.handler.LoggingHandler$Level#
#log.expression         #A SpEL expression (against the#payload
#java.lang.String       #
#      #                  #incoming message) to evaluate #
#
#      #                  #as the logged message.    #
#
#####
```

The application properties for the `time` and `log` apps can be specified at the time of stream creation as follows:

```
dataflow:> stream create --definition "time --fixed-delay=5 / log --level=WARN" --name ticktock
```

Note that the properties `fixed-delay` and `level` defined above for the apps `time` and `log` are the 'short-form' property names provided by the shell completion. These 'short-form' property names are applicable only for the white-listed properties and in all other cases, only *fully qualified* property names should be used.

Passing application properties when deploying a stream

The application properties can also be specified when deploying a stream. When specified during deployment, these application properties can either be specified as 'short-form' property names (applicable for white-listed properties) or *fully qualified* property names. The application properties should have the prefix "app.<appName/label>".

For example, the stream

```
dataflow:> stream create --definition "time / log" --name ticktock
```

can be deployed with application properties using the 'short-form' property names:

```
dataflow:>stream deploy ticktock --properties "app.time.fixed-delay=5,app.log.level=ERROR"
```

When using the app label,

```
stream create ticktock --definition "a: time / b: log"
```

the application properties can be defined as:

```
stream deploy ticktock --properties "app.a.fixed-delay=4,app.b.level=ERROR"
```

Passing stream partition properties during stream deployment

A common pattern in stream processing is to partition the data as it is streamed. This entails deploying multiple instances of a message consuming app and using content-based routing so that messages with a given key (as determined at runtime) are always routed to the same app instance. You can pass the partition properties during stream deployment to declaratively configure a partitioning strategy to route each message to a specific consumer instance.

See below for examples of deploying partitioned streams:

app.[app/label name].producer.partitionKeyExtractorClass

The class name of a PartitionKeyExtractorStrategy (default `null`)

app.[app/label name].producer.partitionKeyExpression

A SpEL expression, evaluated against the message, to determine the partition key; only applies if `partitionKeyExtractorClass` is null. If both are null, the app is not partitioned (default `null`)

app.[app/label name].producer.partitionSelectorClass

The class name of a PartitionSelectorStrategy (default `null`)

app.[app/label name].producer.partitionSelectorExpression

A SpEL expression, evaluated against the partition key, to determine the partition index to which the message will be routed. The final partition index will be the return value (an integer) modulo `[nextModule].count`. If both the class and expression are null, the underlying binder's default PartitionSelectorStrategy will be applied to the key (default `null`)

In summary, an app is partitioned if its count is `> 1` and the previous app has a `partitionKeyExtractorClass` or `partitionKeyExpression` (class takes precedence). When a partition key is extracted, the partitioned app instance is determined by invoking the `partitionSelectorClass`, if present, or the `partitionSelectorExpression` % `partitionCount`, where `partitionCount` is application count in the case of RabbitMQ, and the underlying partition count of the topic in the case of Kafka.

If neither a `partitionSelectorClass` nor a `partitionSelectorExpression` is present the result is `key.hashCode() % partitionCount`.

Overriding application properties during stream deployment

Application properties that are defined during deployment override the same properties defined during the stream creation.

For example, the following stream has application properties defined during stream creation:

```
dataflow:> stream create --definition "time --fixed-delay=5 / log --level=WARN" --name ticktock
```

To override these application properties, one can specify the new property values during deployment:

```
dataflow:>stream deploy ticktock --properties "app.time.fixed-delay=4,app.log.level=ERROR"
```

17.2 Deployment properties

When deploying the stream, properties that control the deployment of the apps into the target platform are known as `deployment` properties. For instance, one can specify how many instances need to be deployed for the specific application defined in the stream using the deployment property called `count`.

Passing instance count as deployment property

If you would like to have multiple instances of an application in the stream, you can include a property with the `deploy` command:

```
dataflow:> stream deploy --name ticktock --properties "app.time.count=3"
```

Note that `count` is the **reserved** property name used by the underlying deployer. Hence, if the application also has a custom property named `count`, it is **not** supported when specified in 'short-form' form during stream *deployment* as it could conflict with the *instance* count deployer property. Instead, the `count` as a custom application property can be specified in its *fully qualified* form (example: `app.foo.bar.count`) during stream *deployment* or it can be specified using 'short-form' or *fully qualified* form during the stream *creation* where it will be considered as an app property.



Important

See [Chapter 24, Using Labels in a Stream](#).

Inline vs file reference properties

When using the Spring Cloud Dataflow Shell, there are two ways to provide deployment properties: either **inline** or via a **file reference**. Those two ways are exclusive and documented below:

Inline properties

use the `--properties` shell option and list properties as a comma separated list of key=value pairs, like so:

```
stream deploy foo
--properties "app.transform.count=2,app.transform.producer.partitionKeyExpression=payload"
```

Using a file reference

use the `--propertiesFile` option and point it to a local Java `.properties` file (i.e. that lives in the filesystem of the machine running the shell). Being read as a `.properties` file, normal rules

apply (ISO 8859-1 encoding, =, <space> or : delimiter, etc.) although we recommend using = as a key-value pair delimiter for consistency:

```
stream deploy foo --propertiesFile myprops.properties
```

where `myprops.properties` contains:

```
app.transform.count=2
app.transform.producer.partitionKeyExpression=payload
```

Both the above properties will be passed as deployment properties for the stream `foo` above.

18. Destroying a Stream

You can delete a stream by issuing the `stream destroy` command from the shell:

```
dataflow:> stream destroy --name ticktock
```

If the stream was deployed, it will be undeployed before the stream definition is deleted.

19. Deploying and Undeploying Streams

Often you will want to stop a stream, but retain the name and definition for future use. In that case you can `undeploy` the stream by name and issue the `deploy` command at a later time to restart it.

```
dataflow:> stream undeploy --name ticktock  
dataflow:> stream deploy --name ticktock
```

20. Other Source and Sink Application Types

Let's try something a bit more complicated and swap out the `time` source for something else. Another supported source type is `http`, which accepts data for ingestion over HTTP POSTs. Note that the `http` source accepts data on a different port from the Data Flow Server (default 8080). By default the port is randomly assigned.

To create a stream using an `http` source, but still using the same `log` sink, we would change the original command above to

```
dataflow:> stream create --definition "http | log" --name myhttpstream --deploy
```

which will produce the following output from the server

```
2016-06-01 09:47:58.920 INFO 79016 --- [io-9393-exec-10] o.s.c.d.spi.local.LocalAppDeployer :
    deploying app myhttpstream.log instance 0
    Logs will be in /var/folders/wn/8jxm_tbdlvj28c8vj37n900m0000gn/T/spring-cloud-
dataflow-912434582726479179/myhttpstream-1464788878747/myhttpstream.log
2016-06-01 09:48:06.396 INFO 79016 --- [io-9393-exec-10] o.s.c.d.spi.local.LocalAppDeployer :
    deploying app myhttpstream.http instance 0
    Logs will be in /var/folders/wn/8jxm_tbdlvj28c8vj37n900m0000gn/T/spring-cloud-
dataflow-912434582726479179/myhttpstream-1464788886383/myhttpstream.http
```

Note that we don't see any other output this time until we actually post some data (using a shell command). In order to see the randomly assigned port on which the `http` source is listening, execute:

```
dataflow:> runtime apps
```

You should see that the corresponding `http` source has a `url` property containing the host and port information on which it is listening. You are now ready to post to that url, e.g.:

```
dataflow:> http post --target http://localhost:1234 --data "hello"
dataflow:> http post --target http://localhost:1234 --data "goodbye"
```

and the stream will then funnel the data from the `http` source to the output log implemented by the `log` sink

```
2016-06-01 09:50:22.121 INFO 79654 --- [ kafka-binder-] log.sink : hello
2016-06-01 09:50:26.810 INFO 79654 --- [ kafka-binder-] log.sink : goodbye
```

Of course, we could also change the sink implementation. You could pipe the output to a file (`file`), to `hadoop` (`hdfs`) or to any of the other sink apps which are available. You can also define your own apps.

21. Simple Stream Processing

As an example of a simple processing step, we can transform the payload of the HTTP posted data to upper case using the stream definitions

```
http | transform --expression=payload.toUpperCase() | log
```

To create this stream enter the following command in the shell

```
dataflow:> stream create --definition "http | transform --expression=payload.toUpperCase() | log" --name  
mystream --deploy
```

Posting some data (using a shell command)

```
dataflow:> http post --target http://localhost:1234 --data "hello"
```

Will result in an uppercased 'HELLO' in the log

```
2016-06-01 09:54:37.749 INFO 80083 --- [ kafka-binder-] log.sink : HELLO
```

22. Stateful Stream Processing

To demonstrate the data partitioning functionality, let's deploy the following stream with Kafka as the binder.

```
dataflow:>stream create --name words --definition "http --server.port=9900 | splitter --
expression=payload.split(' ') | log"
Created new stream 'words'

dataflow:>stream deploy words --properties
"app.splitter.producer.partitionKeyExpression=payload,app.log.count=2"
Deployed stream 'words'

dataflow:>http post --target http://localhost:9900 --data "How much wood would a woodchuck chuck if a
woodchuck could chuck wood"
> POST (text/plain;Charset=UTF-8) http://localhost:9900 How much wood would a woodchuck chuck if a
woodchuck could chuck wood
> 202 ACCEPTED
```

You'll see the following in the server logs.

```
2016-06-05 18:33:24.982 INFO 58039 --- [nio-9393-exec-9] o.s.c.d.spi.local.LocalAppDeployer :
  deploying app words.log instance 0
  Logs will be in /var/folders/c3/ctx7_rns6x30tq7rb76wzqwr0000gp/T/spring-cloud-
dataflow-694182453710731989/words-1465176804970/words.log
2016-06-05 18:33:24.988 INFO 58039 --- [nio-9393-exec-9] o.s.c.d.spi.local.LocalAppDeployer :
  deploying app words.log instance 1
  Logs will be in /var/folders/c3/ctx7_rns6x30tq7rb76wzqwr0000gp/T/spring-cloud-
dataflow-694182453710731989/words-1465176804970/words.log
```

Review the `words.log instance 0` logs:

```
2016-06-05 18:35:47.047 INFO 58638 --- [ kafka-binder-] log.sink : How
2016-06-05 18:35:47.066 INFO 58638 --- [ kafka-binder-] log.sink :
chuck
2016-06-05 18:35:47.066 INFO 58638 --- [ kafka-binder-] log.sink :
```

Review the `words.log instance 1` logs:

```
2016-06-05 18:35:47.047 INFO 58639 --- [ kafka-binder-] log.sink :
much
2016-06-05 18:35:47.066 INFO 58639 --- [ kafka-binder-] log.sink :
wood
2016-06-05 18:35:47.066 INFO 58639 --- [ kafka-binder-] log.sink :
would
2016-06-05 18:35:47.066 INFO 58639 --- [ kafka-binder-] log.sink : a
2016-06-05 18:35:47.066 INFO 58639 --- [ kafka-binder-] log.sink :
woodchuck
2016-06-05 18:35:47.067 INFO 58639 --- [ kafka-binder-] log.sink : if
2016-06-05 18:35:47.067 INFO 58639 --- [ kafka-binder-] log.sink : a
2016-06-05 18:35:47.067 INFO 58639 --- [ kafka-binder-] log.sink :
woodchuck
2016-06-05 18:35:47.067 INFO 58639 --- [ kafka-binder-] log.sink :
could
2016-06-05 18:35:47.067 INFO 58639 --- [ kafka-binder-] log.sink :
wood
```

This shows that payload splits that contain the same word are routed to the same application instance.

23. Tap a Stream

Taps can be created at various producer endpoints in a stream. For a stream like this:

```
stream create --definition "http | step1: transform --expression=payload.toUpperCase() | step2:
transform --expression=payload+'!' | log" --name mainstream --deploy
```

taps can be created at the output of `http`, `step1` and `step2`.

To create a stream that acts as a 'tap' on another stream requires to specify the `source destination name` for the tap stream. The syntax for source destination name is:

```
`:<stream-name>.<label/app-name>`
```

To create a tap at the output of `http` in the stream above, the source destination name is `mainstream.http` To create a tap at the output of the first transform app in the stream above, the source destination name is `mainstream.step1`

The tap stream DSL looks like this:

```
stream create --definition ":mainstream.http > counter" --name tap_at_http --deploy

stream create --definition ":mainstream.step1 > jdbc" --name tap_at_step1_transformer --deploy
```

Note the colon (:) prefix before the destination names. The colon allows the parser to recognize this as a destination name instead of an app name.

24. Using Labels in a Stream

When a stream is comprised of multiple apps with the same name, they must be qualified with labels:

```
stream create --definition "http | firstLabel: transform --expression=payload.toUpperCase() |  
secondLabel: transform --expression=payload+'!' | log" --name myStreamWithLabels --deploy
```

25. Explicit Broker Destinations in a Stream

One can connect to a specific destination name located in the broker (Rabbit, Kafka etc.,) either at the `source` or at the `sink` position.

The following stream has the destination name at the `source` position:

```
stream create --definition ":myDestination > log" --name ingest_from_broker --deploy
```

This stream receives messages from the destination `myDestination` located at the broker and connects it to the `log` app.

The following stream has the destination name at the `sink` position:

```
stream create --definition "http > :myDestination" --name ingest_to_broker --deploy
```

This stream sends the messages from the `http` app to the destination `myDestination` located at the broker.

From the above streams, notice that the `http` and `log` apps are interacting with each other via the broker (through the destination `myDestination`) rather than having a pipe directly between `http` and `log` within a single stream.

It is also possible to connect two different destinations (`source` and `sink` positions) at the broker in a stream.

```
stream create --definition ":destination1 > :destination2" --name bridge_destinations --deploy
```

In the above stream, both the destinations (`destination1` and `destination2`) are located in the broker. The messages flow from the source destination to the sink destination via a `bridge` app that connects them.

26. Directed Graphs in a Stream

If directed graphs are needed instead of the simple linear streams described above, two features are relevant.

First, named destinations may be used as a way to combine the output from multiple streams or for multiple consumers to share the output from a single stream. This can be done using the DSL syntax `http > :mydestination` or `:mydestination > log`.

Second, you may need to determine the output channel of a stream based on some information that is only known at runtime. In that case, a router may be used in the sink position of a stream definition. For more information, refer to the Router Sink starter's [README](#).

26.1 Common application properties

In addition to configuration via DSL, Spring Cloud Data Flow provides a mechanism for setting common properties to all the streaming applications that are launched by it. This can be done by adding properties prefixed with `spring.cloud.dataflow.applicationProperties.stream` when starting the server. When doing so, the server will pass all the properties, without the prefix, to the instances it launches.

For example, all the launched applications can be configured to use a specific Kafka broker by launching the configuration server with the following options:

```
--  
spring.cloud.dataflow.applicationProperties.stream.spring.cloud.stream.kafka.binder.brokers=192.168.1.100:9092  
--  
spring.cloud.dataflow.applicationProperties.stream.spring.cloud.stream.kafka.binder.zkNodes=192.168.1.100:2181
```

This will cause the properties `spring.cloud.stream.kafka.binder.brokers` and `spring.cloud.stream.kafka.binder.zkNodes` to be passed to all the launched applications.



Note

Properties configured using this mechanism have lower precedence than stream deployment properties. They will be overridden if a property with the same key is specified at stream deployment time (e.g. `app.http.spring.cloud.stream.kafka.binder.brokers` will override the common property).

27. Stream applications with multiple binder configurations

In some cases, a stream can have its applications bound to multiple spring cloud stream binders when they are required to connect to different messaging middleware configurations. In those cases, it is important to make sure the applications are configured appropriately with their binder configurations. For example, let's consider the following stream:

```
http | transform --expression=payload.toUpperCase() | log
```

and in this stream, each application connects to messaging middleware in the following way:

```
Http source sends events to RabbitMQ (rabbit1)
Transform processor receives events from RabbitMQ (rabbit1) and sends the processed events into Kafka (kafka1)
Log sink receives events from Kafka (kafka1)
```

Here, `rabbit1` and `kafka1` are the binder names given in the spring cloud stream application properties. Based on this setup, the applications will have the following binder(s) in their classpath with the appropriate configuration:

```
Http - Rabbit binder
Transform - Both Kafka and Rabbit binders
Log - Kafka binder
```

The `spring-cloud-stream` `binder` configuration properties can be set within the applications themselves. If not, they can be passed via `deployment` properties when the stream is deployed.

For example,

```
dataflow:>stream create --definition "http | transform --expression=payload.toUpperCase() | log" --name mystream
```

```
dataflow:>stream deploy mystream --properties
"app.http.spring.cloud.stream.bindings.output.binder=rabbit1,app.transform.spring.cloud.stream.bindings.input.binder=rabbit1,app.transform.spring.cloud.stream.bindings.output.binder=kafka1,app.log.spring.cloud.stream.bindings.input.binder=kafka1"
```

One can override any of the binder configuration properties by specifying them via deployment properties.

Part V. Dashboard

This section describe how to use the Dashboard of Spring Cloud Data Flow.

28. Introduction

Spring Cloud Data Flow provides a browser-based GUI which currently has 6 sections:

- **Apps** Lists all available applications and provides the control to register/unregister them
- **Runtime** Provides the Data Flow cluster view with the list of all running applications
- **Streams** Deploy/undeploy Stream Definitions
- **Tasks** List, create, launch and destroy Task Definitions
- **Jobs** Perform Batch Job related functions
- **Analytics** Create data visualizations for the various analytics applications

Upon starting Spring Cloud Data Flow, the Dashboard is available at:

`http://<host>:<port>/dashboard`

For example: <http://localhost:9393/dashboard>

If you have enabled https, then it will be located at `https://localhost:9393/dashboard`. If you have enabled security, a login form is available at `http://localhost:9393/dashboard/#/login`.

Note: The default Dashboard server port is 9393

About

Spring Cloud Data Flow is a unified, distributed, and extensible system for data ingestion, real time analytics, batch processing, and data export. The project's goal is to simplify the development of big data applications.

Dataflow Server Implementation	
Name	spring-cloud-dataflow-server-local
Version	1.0.0.BUILD-SNAPSHOT (7188a69)
Description	Local Data Flow Server

Need Help or Found an Issue?

Project Page	http://cloud.spring.io/spring-cloud-dataflow/
Sources	https://github.com/spring-cloud/spring-cloud-dataflow
Documentation	http://docs.spring.io/spring-cloud-dataflow/docs/current/reference/html/
API Docs	http://docs.spring.io/spring-cloud-dataflow/docs/current/api/
Support Forum	http://stackoverflow.com/questions/tagged/spring-cloud
Issue Tracker	https://github.com/spring-cloud/spring-cloud-dataflow/issues

Figure 28.1. The Spring Cloud Data Flow Dashboard

29. Apps

The *Apps* section of the Dashboard lists all the available applications and provides the control to register/unregister them (if applicable). By clicking on the magnifying glass, you will get a listing of available definition properties.

The screenshot displays the 'Apps' section of the Spring Cloud Data Flow Server dashboard. At the top, there is a navigation bar with the 'spring' logo and tabs for APPS, RUNTIME, STREAMS, TASKS, JOBS, ANALYTICS, and ABOUT. The 'APPS' tab is currently selected. Below the navigation bar, the 'Apps' section is titled, followed by a description: 'This section lists all the available applications and provides the control to register/unregister them (if applicable).'.

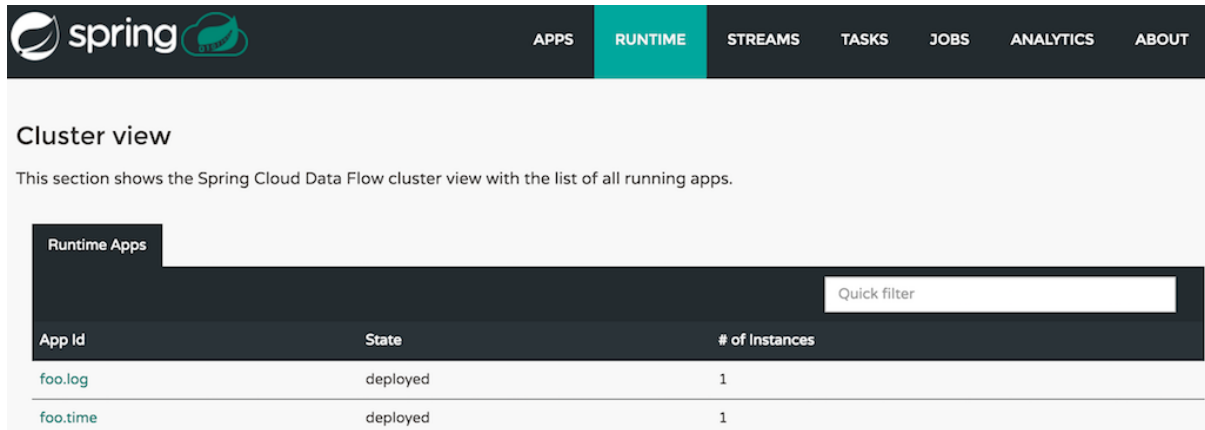
Below the description, there is a section titled 'All Applications'. It contains two buttons: '+ Register Application(s)' and 'Unregister Application(s)'. To the right of these buttons is a 'Quick filter' input field. Below this is a table listing the available applications.

Name	Type	URI	Actions
<input type="checkbox"/> file	source	maven://org.springframework.cloud.stream.app:file-source-kafka:1.0.0.BUILD-SNAPSHOT	<input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/> ftp	source	maven://org.springframework.cloud.stream.app:ftp-source-kafka:1.0.0.BUILD-SNAPSHOT	<input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/> http	source	maven://org.springframework.cloud.stream.app:http-source-kafka:1.0.0.BUILD-SNAPSHOT	<input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/> jdbc	source	maven://org.springframework.cloud.stream.app:jdbc-source-kafka:1.0.0.BUILD-SNAPSHOT	<input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/> jms	source	maven://org.springframework.cloud.stream.app:jms-source-kafka:1.0.0.BUILD-SNAPSHOT	<input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/> load-generator	source	maven://org.springframework.cloud.stream.app:load-generator-source-kafka:1.0.0.BUILD-SNAPSHOT	<input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/> rabbit	source	maven://org.springframework.cloud.stream.app:rabbit-source-kafka:1.0.0.BUILD-SNAPSHOT	<input type="checkbox"/> <input type="checkbox"/>
<input type="checkbox"/> sftp	source	maven://org.springframework.cloud.stream.app:sftp-source-kafka:1.0.0.BUILD-SNAPSHOT	<input type="checkbox"/> <input type="checkbox"/>

Figure 29.1. List of Available Applications

30. Runtime

The *Runtime* section of the Dashboard application shows the Spring Cloud Data Flow cluster view with the list of all running applications. For each runtime app the state of the deployment and the number of deployed instances is shown. A list of the used deployment properties is available by clicking on the app id.



The screenshot shows the 'Runtime' tab selected in the dashboard. Below the navigation bar, there's a 'Cluster view' section with a description. A 'Runtime Apps' tab is active, displaying a table of running applications. The table has columns for 'App Id', 'State', and '# of Instances'. Two applications are listed: 'foo.log' and 'foo.time', both in a 'deployed' state with 1 instance each. A 'Quick filter' input field is located to the right of the table header.

App Id	State	# of Instances
foo.log	deployed	1
foo.time	deployed	1

Figure 30.1. List of Running Applications

31. Streams

The *Streams* section of the Dashboard provides the *Definitions* tab that provides a listing of Stream definitions. There you have the option to **deploy** or **undeploy** those stream definitions. Additionally you can remove the definition by clicking on **destroy**.

The screenshot shows the 'Streams' section of the Spring Cloud Data Flow Dashboard. The top navigation bar includes 'APPS', 'RUNTIME', 'STREAMS' (highlighted), 'TASKS', 'JOBS', 'ANALYTICS', and 'ABOUT'. Below the navigation bar, the 'Streams' section is titled, followed by a description: 'This section lists all the stream definitions and provides the ability to deploy/undeploy or destroy streams.' A 'Definitions' tab is active, and a 'Create Stream' link is visible. A table lists stream definitions with columns for Name, Definition, Status, and Actions. A 'Quick filter' input field is present above the table. The table contains one entry: 'foo' with definition 'time | log' and status 'deployed'. The Actions column for this entry contains three buttons: 'Undeploy', 'Deploy', and 'Destroy'.

Name	Definition	Status	Actions
foo	time log	deployed	<button>Undeploy</button> <button>Deploy</button> <button>Destroy</button>

Figure 31.1. List of Stream Definitions

32. Create Stream

The *Create Stream* section of the Dashboard includes the [Spring Flo](#) designer tab that provides the canvas application, offering an interactive graphical interface for creating data pipelines.

In this tab, you can:

- Create, manage, and visualize stream pipelines using DSL, a graphical canvas, or both
- Write pipelines via DSL with content-assist and auto-complete
- Use auto-adjustment and grid-layout capabilities in the GUI for simpler and interactive organization of pipelines

Watch this [screencast](#) that highlights some of the "Flo for Spring Cloud Data Flow" capabilities. Spring Flo [wiki](#) includes more detailed content on core Flo capabilities.

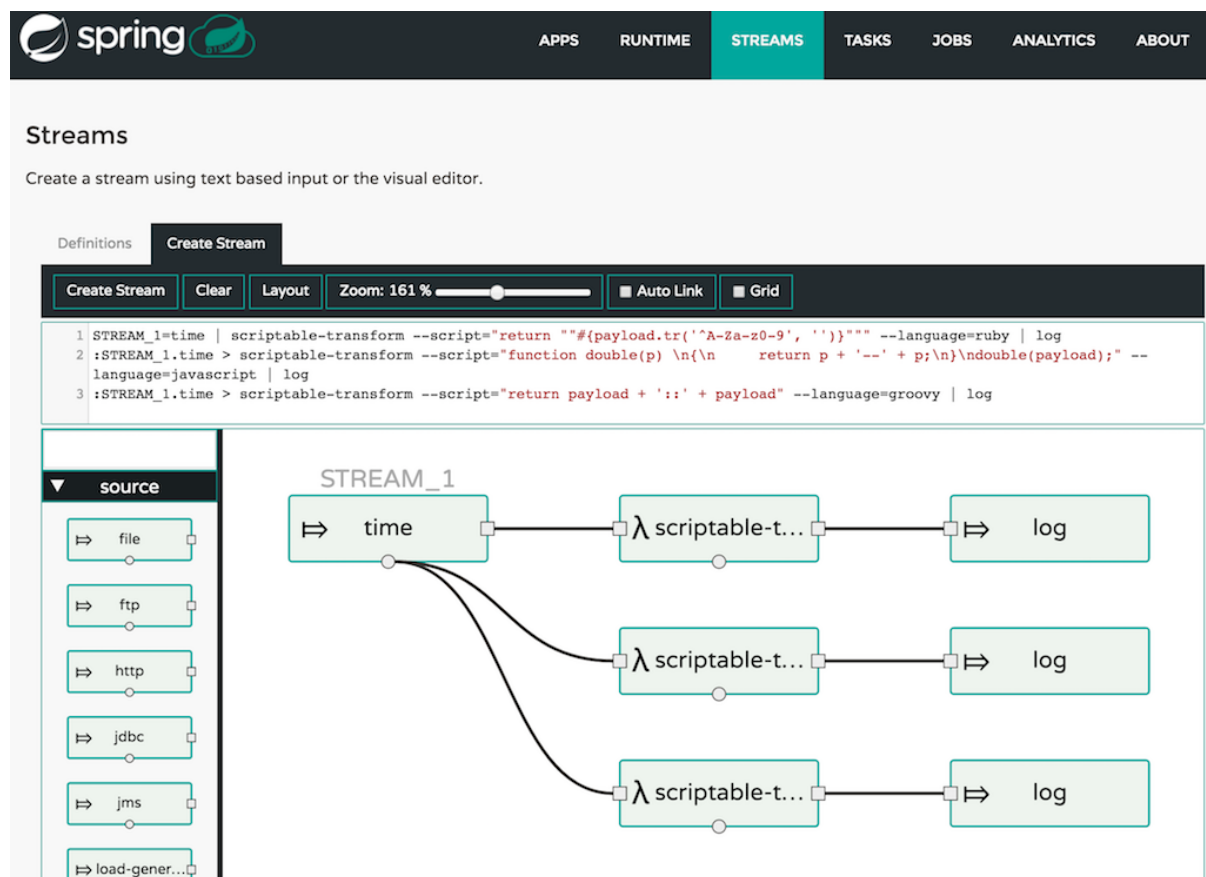


Figure 32.1. Flo for Spring Cloud Data Flow

33. Tasks

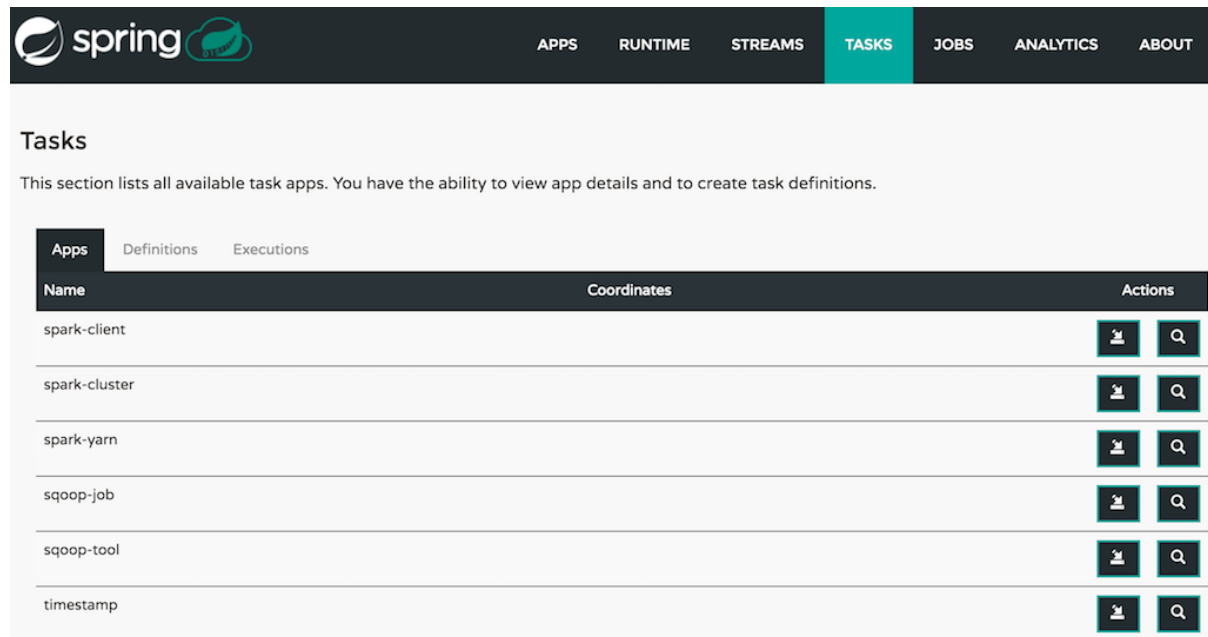
The *Tasks* section of the Dashboard currently has three tabs:

- Apps
- Definitions
- Executions

33.1 Apps

Apps encapsulate a unit of work into a reusable component. Within the Data Flow runtime environment *Apps* allow users to create definitions for *Streams* as well as *Tasks*. Consequently, the *Apps* tab within the *Tasks* section allows users to create *Task* definitions.

Note: You will also use this tab to create Batch Jobs.















Name	Coordinates	Actions
spark-client		 
spark-cluster		 
spark-yarn		 
sqoop-job		 
sqoop-tool		 
timestamp		 

Figure 33.1. List of Task Apps

On this screen you can perform the following actions:

- View details such as the task app options.
- Create a Task Definition from the respective App.

Create a Task Definition from a selected Task App

On this screen you can create a new Task Definition. As a minimum you must provide a name for the new definition. You will also have the option to specify various properties that are used during the deployment of the app.

Note: Each parameter is only included if the *Include* checkbox is selected.

View Task App Details

On this page you can view the details of a selected task app, including the list of available options (properties) for that app.

33.2 Definitions

This page lists the Data Flow Task definitions and provides actions to **launch** or **destroy** those tasks.

The screenshot shows the 'Tasks' page with the 'Definitions' tab selected. The page header includes the Spring logo and navigation links: APPS, RUNTIME, STREAMS, TASKS (active), JOBS, ANALYTICS, and ABOUT. Below the header, the 'Tasks' section is titled 'Tasks' with a description: 'This section lists all the task definitions and allows you to create, launch and destroy them.' There are three tabs: 'Apps', 'Definitions' (active), and 'Executions'. A 'Quick filter' input field is present. Below the filter is a table with columns: Name, Definition, Status, and Actions. The table contains one row for 'demo-timestamp' with definition 'timestamp' and status 'complete'. The 'Actions' column for this row contains two buttons: 'Launch' and 'Destroy'.

Name	Definition	Status	Actions
demo-timestamp	timestamp	complete	▶ Launch ✕ Destroy

Figure 33.2. List of Task Definitions

Launching Tasks

Once the task definition is created, they can be launched through the Dashboard as well. Navigate to the **Definitions** tab. Select the Task you want to launch by pressing Launch.

On the following screen, you can define one or more Task parameters by entering:

- Parameter Key
- Parameter Value

Task parameters are not typed.

33.3 Executions

The screenshot shows the 'Tasks' page with the 'Executions' tab selected. The page header is the same as Figure 33.2. Below the header, the 'Tasks' section is titled 'Tasks' with a description: 'This section lists all the available task executions.' There are three tabs: 'Apps', 'Definitions', and 'Executions' (active). A 'Quick filter' input field is present. Below the filter is a table with columns: Execution Id, Task Name, Start Time, End Time, and Exit Code. The table contains three rows of execution data.

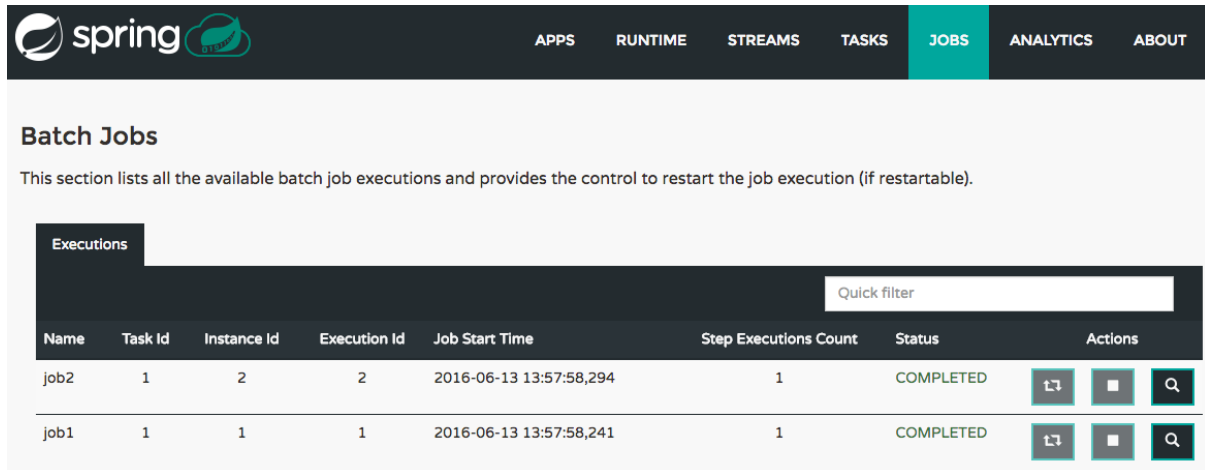
Execution Id	Task Name	Start Time	End Time	Exit Code
3	demo-timestamp	2016-06-11 14:28:10,900	2016-06-11 14:28:10,931	0
2	demo-timestamp	2016-06-11 14:28:09,216	2016-06-11 14:28:09,257	0
1	demo-timestamp	2016-06-11 14:27:13,113	2016-06-11 14:27:13,145	0

Figure 33.3. List of Task Executions

34. Jobs

The *Jobs* section of the Dashboard allows you to inspect **Batch Jobs**. The main section of the screen provides a list of Job Executions. **Batch Jobs** are **Tasks** that were executing one or more **Batch Job**. As such each Job Execution has a back reference to the **Task Execution Id** (Task Id).

In case of a failed job, you can also restart the task. When dealing with long-running Batch Jobs, you can also request to stop it.



Name	Task Id	Instance Id	Execution Id	Job Start Time	Step Executions Count	Status	Actions
job2	1	2	2	2016-06-13 13:57:58,294	1	COMPLETED	[Restart] [Stop] [Search]
job1	1	1	1	2016-06-13 13:57:58,241	1	COMPLETED	[Restart] [Stop] [Search]

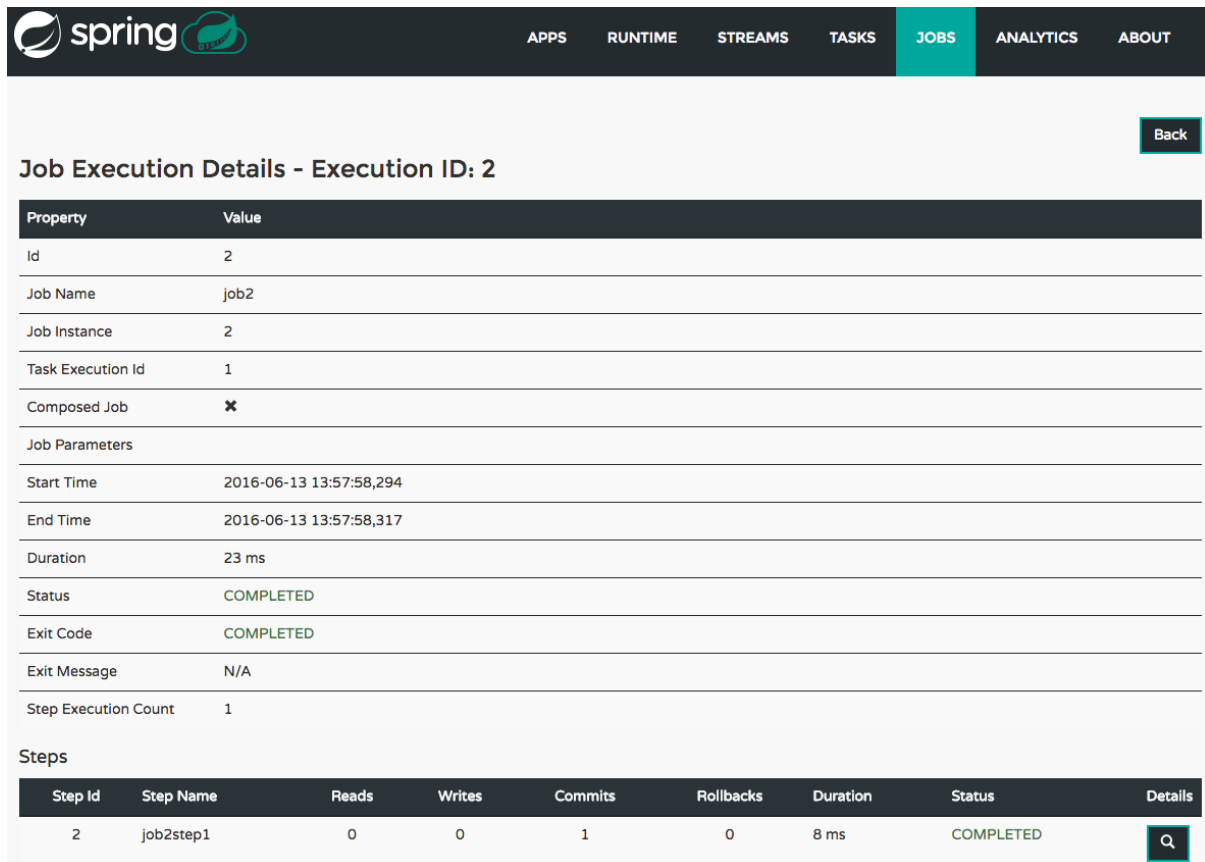
Figure 34.1. List of Job Executions

34.1 List job executions

This page lists the Batch Job Executions and provides the option to **restart** or **stop** a specific job execution, provided the operation is available. Furthermore, you have the option to view the Job execution details.

The list of Job Executions also shows the state of the underlying Job Definition. Thus, if the underlying definition has been deleted, *deleted* will be shown.

Job execution details



Job Execution Details - Execution ID: 2

Property	Value
Id	2
Job Name	job2
Job Instance	2
Task Execution Id	1
Composed Job	✖
Job Parameters	
Start Time	2016-06-13 13:57:58,294
End Time	2016-06-13 13:57:58,317
Duration	23 ms
Status	COMPLETED
Exit Code	COMPLETED
Exit Message	N/A
Step Execution Count	1

Steps


Step Id	Step Name	Reads	Writes	Commits	Rollbacks	Duration	Status	Details
2	job2step1	0	0	1	0	8 ms	COMPLETED	

Figure 34.2. Job Execution Details

The Job Execution Details screen also contains a list of the executed steps. You can further drill into the *Step Execution Details* by clicking onto the magnifying glass.

Step execution details

On the top of the page, you will see progress indicator the respective step, with the option to refresh the indicator. Furthermore, a link is provided to view the *step execution history*.

The Step Execution details screen provides a complete list of all Step Execution Context key/value pairs.



Important

In case of exceptions, the *Exit Description* field will contain additional error information. Please be aware, though, that this field can only have a maximum of **2500 characters**. Therefore, in case of long exception stacktraces, trimming of error messages may occur. In that case, please refer to the server log files for further details.

Step Execution Progress

On this screen, you can see a progress bar indicator in regards to the execution of the current step. Under the **Step Execution History**, you can also view various metrics associated with the selected step such as **duration**, **read counts**, **write counts** etc.

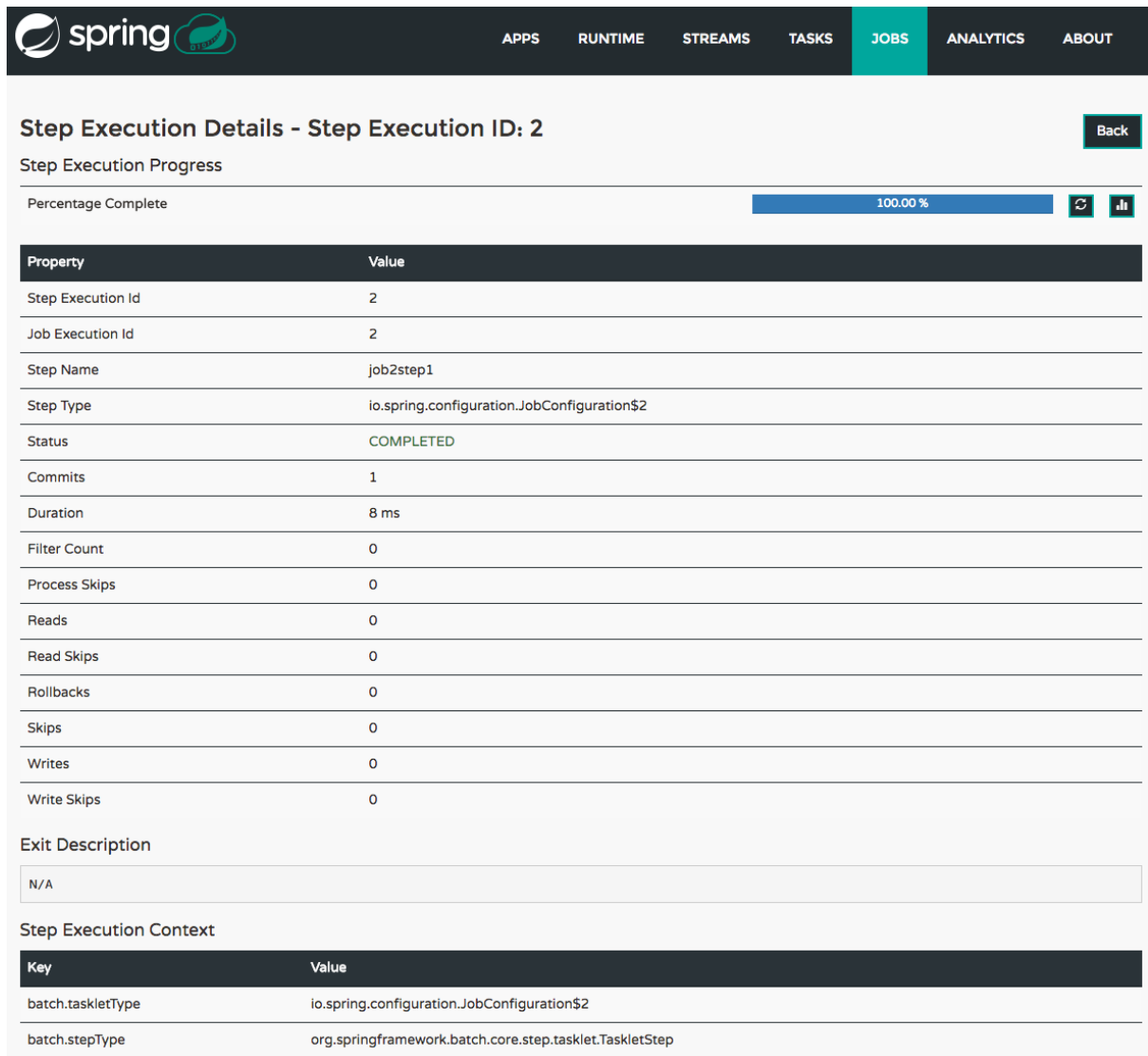


Figure 34.3. Step Execution History

35. Analytics

The *Analytics* section of the Dashboard provided data visualization capabilities for the various analytics applications available in *Spring Cloud Data Flow*:

- Counters
- Field-Value Counters

For example, if you have created the `springtweets` stream and the corresponding counter in the [Counter chapter](#), you can now easily create the corresponding graph from within the **Dashboard** tab:

1. Under `Metric Type`, select `Counters` from the select box
2. Under `Stream`, select `tweetcount`
3. Under `Visualization`, select the desired chart option, `Bar Chart`

Using the icons to the right, you can add additional charts to the Dashboard, re-arrange the order of created dashboards or remove data visualizations.

Part VI. Appendices

Appendix A. Migrating from Spring XD to Spring Cloud Data Flow

A.1 Terminology Changes

Old	New
XD-Admin	Server (<i>implementations</i> : local, cloud foundry, apache yarn, kubernetes, and apache mesos)
XD-Container	N/A
Modules	Applications
Admin UI	Dashboard
Message Bus	Binders
Batch / Job	Task

A.2 Modules to Applications

If you have custom Spring XD modules, you'd have to refactor them to use Spring Cloud Stream and Spring Cloud Task annotations, with updated dependencies and built as normal Spring Boot "applications".

Custom Applications

- Spring XD's stream and batch modules are refactored into [Spring Cloud Stream](#) and [Spring Cloud Task](#) application-starters, respectively. These applications can be used as the reference while refactoring Spring XD modules
- There are also some samples for [Stream](#) and [Task](#) applications for reference
- If you'd like to create a brand new custom application, use the getting started guide for [Stream](#) and [Task](#) applications and as well as review the development [guide](#)
- Alternatively, if you'd like to patch any of the out-of-the-box stream applications, you can follow the procedure [here](#)

Application Registration

- Custom Stream/Task application requires being installed to a maven repository for Local, YARN, and CF implementations or as docker images, when deploying to Kubernetes and Mesos. Other than maven and docker resolution, you can also resolve application artifacts from `http`, `file`, or as `hdfs` coordinates
- Unlike Spring XD, you do not have to upload the application bits while registering custom applications anymore; instead, you're expected to [register](#) the application coordinates that are hosted in the maven repository or by other means as discussed in the previous bullet

- By default, none of the out-of-the-box applications are preloaded already. It is intentionally designed to provide the flexibility to register app(s), as you find appropriate for the given use-case requirement
- Depending on the binder choice, you can manually add the appropriate binder dependency to build applications specific to that binder-type. Alternatively, you can follow the Spring Initializr [procedure](#) to create an application with binder embedded in it

Application Properties

- counter-sink:
 - The peripheral `redis` is not required in Spring Cloud Data Flow. If you intend to use the `counter-sink`, then `redis` becomes required, and you're expected to have your own running `redis` cluster
- field-value-counter-sink:
 - The peripheral `redis` is not required in Spring Cloud Data Flow. If you intend to use the `field-value-counter-sink`, then `redis` becomes required, and you're expected to have your own running `redis` cluster
- aggregate-counter-sink:
 - The peripheral `redis` is not required in Spring Cloud Data Flow. If you intend to use the `aggregate-counter-sink`, then `redis` becomes required, and you're expected to have your own running `redis` cluster

A.3 Message Bus to Binders

Terminology wise, in Spring Cloud Data Flow, the message bus implementation is commonly referred to as binders.

Message Bus

Similar to Spring XD, there's an abstraction available to extend the binder interface. By default, we take the opinionated view of [Apache Kafka](#) and [RabbitMQ](#) as the production-ready binders and are available as GA releases. We also have an experimental version of the [Gemfire](#) binder.

Binders

Selecting a binder is as simple as providing the right binder dependency in the classpath. If you're to choose Kafka as the binder, you'd register stream applications that are pre-built with Kafka binder in it. If you were to create a custom application with Kafka binder, you'd add the following dependency in the classpath.

```
<dependency>
  <groupId>org.springframework.cloud</groupId>
  <artifactId>spring-cloud-stream-binder-kafka</artifactId>
  <version>1.0.2.RELEASE</version>
</dependency>
```

- Spring Cloud Stream supports [Apache Kafka](#), [RabbitMQ](#) and an experimental [Gemfire](#) binder implementation. All binder implementations are maintained and managed in their individual repositories

- Every Stream/Task application can be built with a binder implementation of your choice. All the out-of-the-box applications are pre-built for both Kafka and Rabbit and they're readily available for use as maven artifacts [[stream](#) / [task](#)] or docker images [[stream](#) / [task](#)] Changing the binder requires selecting the right binder [dependency](#). Alternatively, you can download the pre-built application from this version of [Spring Initializr](#) with the desired “binder-starter” dependency

Named Channels

Fundamentally, all the messaging channels are backed by pub/sub semantics. Unlike Spring XD, the messaging channels are backed only by `topics` or `topic-exchange` and there's no representation of `queues` in the new architecture.

- `${xd.module.index}` is not supported anymore; instead, you can directly interact with named destinations
- `stream.index` changes to `:<stream-name>.<label/app-name>`
 - *for instance:* `ticktock.0` changes to `:ticktock.time`
- “topic/queue” prefixes are not required to interact with named-channels
 - *for instance:* `topic:foo` changes to `:foo`
 - *for instance:* `stream create stream1 --definition ":foo > log"`

Directed Graphs

If you're building non-linear streams, you could take advantage of named destinations to build directed graphs.

for instance, in Spring XD:

```
stream create f --definition "queue:foo > transform --expression=payload+'-foo' | log" --deploy
stream create b --definition "queue:bar > transform --expression=payload+'-bar' | log" --deploy
stream create r --definition "http | router --expression=payload.contains('a')?'queue:foo':'queue:bar'"
--deploy
```

for instance, in Spring Cloud Data Flow:

```
stream create f --definition ":foo > transform --expression=payload+'-foo' | log" --deploy
stream create b --definition ":bar > transform --expression=payload+'-bar' | log" --deploy
stream create r --definition "http | router --expression=payload.contains('a')?'foo':'bar'" --deploy
```

A.4 Batch to Tasks

A Task by definition, is any application that does not run forever, including Spring Batch jobs, and they end/stop at some point. Task applications can be majorly used for on-demand use-cases such as database migration, machine learning, scheduled operations etc. Using [Spring Cloud Task](#), users can build Spring Batch jobs as microservice applications.

- Spring Batch [jobs](#) from Spring XD are being refactored to Spring Boot applications a.k.a link: [Spring Cloud Task applications](#)
- Unlike Spring XD, these “Tasks” don't require explicit deployment; instead, a task is ready to be launched directly once the definition is declared

A.5 Shell/DSL Commands

Old Command	New Command
module upload	app register / app import
module list	app list
module info	app info
admin config server	dataflow config server
job create	task create
job launch	task launch
job list	task list
job status	task status
job display	task display
job destroy	task destroy
job execution list	task execution list
runtime modules	runtime apps

A.6 REST-API

Old API	New API
/modules	/apps
/runtime/modules	/runtime/apps
/runtime/modules/{moduleId}	/runtime/apps/{appId}
/jobs/definitions	/task/definitions
/jobs/deployments	/task/deployments

A.7 UI / Flo

The Admin-UI is now renamed as Dashboard. The URI for accessing the Dashboard is changed from localhost:9393/admin-ui to localhost:9393/dashboard

- (New) Apps: Lists all the registered applications that are available for use. This view includes informational details such as the URI and the properties supported by each application. You can also register/unregister applications from this view
- Runtime: Container changes to Runtime. The notion of `xd-container` is gone, replaced by out-of-the-box applications running as autonomous Spring Boot applications. The Runtime tab displays the applications running in the runtime platforms (*implementations*: cloud foundry, apache yarn, apache mesos, or kubernetes). You can click on each application to review relevant details about the application such as where it is running with, and what resources etc.

- [Spring Flo](#) is now an OSS product. Flo for Spring Cloud Data Flow's "Create Stream", the designer-tab comes pre-built in the Dashboard
- (New) Tasks:
 - The sub-tab "Modules" is renamed to "Apps"
 - The sub-tab "Definitions" lists all the Task definitions, including Spring Batch jobs that are orchestrated as Tasks
 - The sub-tab "Executions" lists all the Task execution details similar to Spring XD's Job executions

A.8 Architecture Components

Spring Cloud Data Flow comes with a significantly simplified architecture. In fact, when compared with Spring XD, there are less peripherals that are necessary to operationalize Spring Cloud Data Flow.

ZooKeeper

ZooKeeper is not used in the new architecture.

RDBMS

Spring Cloud Data Flow uses an RDBMS instead of Redis for stream/task definitions, application registration, and for job repositories. The default configuration uses an embedded H2 instance, but Oracle, SqlServer, MySQL/MariaDB, PostgreSQL, H2, and HSQLDB databases are supported. To use Oracle and SqlServer you will need to create your own Data Flow Server using [Spring Initializr](#) and add the appropriate JDBC driver dependency.

Redis

Running a Redis cluster is only required for analytics functionality. Specifically, when the `counter-sink`, `field-value-counter-sink`, or `aggregate-counter-sink` applications are used, it is expected to also have a running instance of Redis cluster.

Cluster Topology

Spring XD's `xd-admin` and `xd-container` server components are replaced by stream and task applications themselves running as autonomous Spring Boot applications. The applications run natively on various platforms including Cloud Foundry, Apache YARN, Apache Mesos, or Kubernetes. You can develop, test, deploy, scale +/-, and interact with (Spring Boot) applications individually, and they can evolve in isolation.

A.9 Central Configuration

To support centralized and consistent management of an application's configuration properties, [Spring Cloud Config](#) client libraries have been included into the Spring Cloud Data Flow server as well as the Spring Cloud Stream applications provided by the Spring Cloud Stream App Starters. You can also [pass common application properties](#) to all streams when the Data Flow Server starts.

A.10 Distribution

Spring Cloud Data Flow is a Spring Boot application. Depending on the platform of your choice, you can download the respective release uber-jar and deploy/push it to the runtime platform (cloud foundry,

apache yarn, kubernetes, or apache mesos). For example, if you're running Spring Cloud Data Flow on Cloud Foundry, you'd download the Cloud Foundry server implementation and do a `cf push` as explained in the [reference guide](#).

A.11 Hadoop Distribution Compatibility

The `hdfs-sink` application builds upon Spring Hadoop 2.4.0 release, so this application is compatible with following Hadoop distributions.

- Cloudera - cdh5
- Pivotal Hadoop - phd30
- Hortonworks Hadoop - hdp24
- Hortonworks Hadoop - hdp23
- Vanilla Hadoop - hadoop26
- Vanilla Hadoop - 2.7.x (default)

A.12 YARN Deployment

Spring Cloud Data Flow can be deployed and used with Apache YARN in two different ways.

- Deploy the server [directly](#) in a YARN cluster
- Leverage Apache Ambari [plugin to provision](#) Spring Cloud Data Flow as a service

A.13 Use Case Comparison

Let's review some use-cases to compare and contrast the differences between Spring XD and Spring Cloud Data Flow.

Use Case #1

(It is assumed both XD and SCDF distributions are already downloaded)

Description: Simple `ticktock` example using local/singlenode.

Spring XD	Spring Cloud Data Flow
Start <code>xd-singlenode</code> server from CLI <pre># xd-singlenode</pre>	Start a binder of your choice Start <code>local-server</code> implementation of SCDF from the CLI <pre># java -jar spring-cloud-dataflow-server-local-1.0.0.BUILD-SNAPSHOT.jar</pre>
Start <code>xd-shell</code> server from the CLI <pre># xd-shell</pre>	Start <code>dataflow-shell</code> server from the CLI

Spring XD	Spring Cloud Data Flow
	<pre># java -jar spring-cloud-dataflow-shell-1.0.0.BUILD-SNAPSHOT.jar</pre>
Create ticktock stream <code>xd:>stream create ticktock --definition "time log" --deploy</code>	Create ticktock stream <code>dataflow:>stream create ticktock --definition "time log" --deploy</code>
Review ticktock results in the xd-singlenode server console	Review ticktock results by tailing the ticktock.log/stdout_log application logs

Use Case #2

(It is assumed both XD and SCDF distributions are already downloaded)

Description: Stream with custom module/application.

Spring XD	Spring Cloud Data Flow
Start xd-singlenode server from CLI <pre># xd-singlenode</pre>	Start a binder of your choice Start local-server implementation of SCDF from the CLI <pre># java -jar spring-cloud-dataflow-server-local-1.0.0.BUILD-SNAPSHOT.jar</pre>
Start xd-shell server from the CLI <pre># xd-shell</pre>	Start dataflow-shell server from the CLI <pre># java -jar spring-cloud-dataflow-shell-1.0.0.BUILD-SNAPSHOT.jar</pre>
Register custom "processor" module to transform payload to a desired format <code>xd:>module upload --name toupper --type processor --file <CUSTOM_JAR_FILE_LOCATION></code>	Register custom "processor" application to transform payload to a desired format <code>dataflow:>app register --name toupper --type processor --uri <MAVEN_URI_COORDINATES></code>
Create a stream with custom module <code>xd:>stream create testupper --definition "http toupper log" --deploy</code>	Create a stream with custom application <code>dataflow:>stream create testupper --definition "http toupper log" --deploy</code>
Review results in the xd-singlenode server console	Review results by tailing the testupper.log/stdout_log application logs

Use Case #3

(It is assumed both XD and SCDF distributions are already downloaded)

Description: Simple batch-job.

Spring XD	Spring Cloud Data Flow
Start xd-singlenode server from CLI <pre># xd-singlenode</pre>	Start local-server implementation of SCDF from the CLI <pre># java -jar spring-cloud-dataflow-server-local-1.0.0.BUILD-SNAPSHOT.jar</pre>
Start xd-shell server from the CLI <pre># xd-shell</pre>	Start dataflow-shell server from the CLI <pre># java -jar spring-cloud-dataflow-shell-1.0.0.BUILD-SNAPSHOT.jar</pre>
Register custom “batch-job” module <pre>xd:>module upload --name simple-batch --type job --file <CUSTOM_JAR_FILE_LOCATION></pre>	Register custom “batch-job” as task application <pre>dataflow:>app register --name simple-batch --type task --uri <MAVEN_URI_COORDINATES></pre>
Create a job with custom batch-job module <pre>xd:>job create batchtest --definition "simple-batch"</pre>	Create a task with custom batch-job application <pre>dataflow:>task create batchtest --definition "simple-batch"</pre>
Deploy job <pre>xd:>job deploy batchtest</pre>	NA
Launch job <pre>xd:>job launch batchtest</pre>	Launch task <pre>dataflow:>task launch batchtest</pre>
Review results in the xd-singlenode server console as well as Jobs tab in UI (executions sub-tab should include all step details)	Review results by tailing the batchtest/ stdout_log application logs as well as Task tab in UI (executions sub-tab should include all step details)

Appendix B. Test Cluster

Here are brief setup instructions for setting up a local Vagrant DC/OS Mesos cluster. The DC/OS Dashboard endpoint will be <http://m1.dcos>, the Mesos endpoint will be <http://m1.dcos/mesos>, the Marathon endpoint will be <http://m1.dcos/marathon> and the Chronos endpoint will be <http://m1.dcos/service/chronos>.

Detailed installation instructions are here: <https://github.com/dcos/dcos-vagrant>

B.1 Clone and Download files

First clone the `dcos-vagrant` project:

```
git clone https://github.com/dcos/dcos-vagrant.git
```

Download the DC/OS version you want to use, we used version 1.8.1 from <https://dcos.io/releases/1.8.1/>, and place this `dcos_generate_config.sh` file in the root directory of the cloned `dcos-project`.

Copy the config example file. From the root directory of the `dcos-vagrant` project execute:

```
cp VagrantConfig.yaml.example VagrantConfig.yaml
```

Update the `etc/config-1.8.yaml` file. Add a line with

```
oauth_enabled: 'false'
```

to the end of this file and save it. This will disable any security for the cluster.

B.2 Start Vagrant VMs for the Mesos cluster

You need to install Vagrant and VirtualBox. We used the DC/OS version 1.8.1 with Vagrant version 1.8.4.

Install Vagrant Host Manager Plugin if it isn't already installed:

```
vagrant plugin install vagrant-hostmanager
```

Before we start the cluster we should define the config file we want to use:

```
export DCOS_CONFIG_PATH=etc/config-1.8.yaml
```

Now we are ready to start the Vagrant VMs. You can select different cluster configurations depending on your needs. We have tested using a single master, three private agents and one bootstrap node using this command:

```
vagrant up m1 a1 a2 a3 boot
```

When you are done with the Mesos cluster you can destroy it using:

```
vagrant destroy -f
```

B.3 Install Chronos as a Service running in Marathon

You can install Chronos by using the following `dcos` command. This requires that you install the [DC/OS CLI](#).

```
dcos package install chronos
```

Appendix C. Building

To build the source you will need to install JDK 1.7.

The build uses the Maven wrapper so you don't have to install a specific version of Maven. To enable the tests for Redis you should run the server before building. See below for more information on how to run Redis.

The main build command is

```
$ ./mvnw clean install
```

You can also add '-DskipTests' if you like, to avoid running the tests.



Note

You can also install Maven ($\geq 3.3.3$) yourself and run the `mvn` command in place of `./mvnw` in the examples below. If you do that you also might need to add `-P spring` if your local Maven settings do not contain repository declarations for Spring pre-release artifacts.



Note

Be aware that you might need to increase the amount of memory available to Maven by setting a `MAVEN_OPTS` environment variable with a value like `-Xmx512m -XX:MaxPermSize=128m`. We try to cover this in the `.mvn` configuration, so if you find you have to do it to make a build succeed, please raise a ticket to get the settings added to source control.

The projects that require middleware generally include a `docker-compose.yml`, so consider using [Docker Compose](#) to run the middleware servers in Docker containers. See the README in the [scripts demo repository](#) for specific instructions about the common cases of mongo, rabbit and redis.

C.1 Documentation

There is a "full" profile that will generate documentation. You can build just the documentation by executing

```
$ ./mvnw clean package -DskipTests -P full -pl spring-cloud-dataflow-server-mesos-docs -am
```

C.2 Working with the code

If you don't have an IDE preference we would recommend that you use [Spring Tools Suite](#) or [Eclipse](#) when working with the code. We use the [m2eclipse](#) eclipse plugin for maven support. Other IDEs and tools should also work without issue.

Importing into eclipse with m2eclipse

We recommend the [m2eclipse](#) eclipse plugin when working with eclipse. If you don't already have m2eclipse installed it is available from the "eclipse marketplace".

Unfortunately m2e does not yet support Maven 3.3, so once the projects are imported into Eclipse you will also need to tell m2eclipse to use the `.settings.xml` file for the projects. If you do not do this

you may see many different errors related to the POMs in the projects. Open your Eclipse preferences, expand the Maven preferences, and select User Settings. In the User Settings field click Browse and navigate to the Spring Cloud project you imported selecting the `.settings.xml` file in that project. Click Apply and then OK to save the preference changes.

**Note**

Alternatively you can copy the repository settings from [.settings.xml](#) into your own `~/ .m2/ settings.xml`.

Importing into eclipse without m2eclipse

If you prefer not to use m2eclipse you can generate eclipse project metadata using the following command:

```
$ ./mvnw eclipse:eclipse
```

The generated eclipse projects can be imported by selecting `import existing projects` from the `file` menu.

Appendix D. Contributing

Spring Cloud is released under the non-restrictive Apache 2.0 license, and follows a very standard Github development process, using Github tracker for issues and merging pull requests into master. If you want to contribute even something trivial please do not hesitate, but follow the guidelines below.

D.1 Sign the Contributor License Agreement

Before we accept a non-trivial patch or pull request we will need you to sign the [contributor's agreement](#). Signing the contributor's agreement does not grant anyone commit rights to the main repository, but it does mean that we can accept your contributions, and you will get an author credit if we do. Active contributors might be asked to join the core team, and given the ability to merge pull requests.

D.2 Code Conventions and Housekeeping

None of these is essential for a pull request, but they will all help. They can also be added after the original pull request but before a merge.

- Use the Spring Framework code format conventions. If you use Eclipse you can import formatter settings using the `eclipse-code-formatter.xml` file from the [Spring Cloud Build](#) project. If using IntelliJ, you can use the [Eclipse Code Formatter Plugin](#) to import the same file.
- Make sure all new `.java` files to have a simple Javadoc class comment with at least an `@author` tag identifying you, and preferably at least a paragraph on what the class is for.
- Add the ASF license header comment to all new `.java` files (copy from existing files in the project)
- Add yourself as an `@author` to the `.java` files that you modify substantially (more than cosmetic changes).
- Add some Javadocs and, if you change the namespace, some XSD doc elements.
- A few unit tests would help a lot as well — someone has to do it.
- If no-one else is using your branch, please rebase it against the current master (or other target branch in the main project).
- When writing a commit message please follow [these conventions](#), if you are fixing an existing issue please add `Fixes gh-XXXX` at the end of the commit message (where XXXX is the issue number).