

CONTINUUNITY APPFABRIC DEVELOPER GUIDE

VERSION 1.4.0, RELEASED MARCH 6, 2013

TABLE OF CONTENTS

1. Introduction	3
1.1. What is the Continuity AppFabric?	3
1.2. What is the Developer Suite?	4
What's in the Box?	4
1.3. How to build Apps using Continuity	4
2. Hello World	5
3. Understanding the Continuity AppFabric	7
3.1. Collect with Streams	8
3.2. Process with Flows.....	8
3.3. Store with Datasets.....	8
Types of Datasets	8
3.4. Query with Procedures	8
3.5. Package with Applications	9
3.6. AppFabric Runtime Editions	9
In-Memory AppFabric	9
Local AppFabric	9
Developer Sandbox	9
Distributed AppFabric.....	9
4. Getting Started with the Local AppFabric	10
4.1. Prerequisites.....	10
4.2. Unpacking the Developer Suite	10
4.3. Building Example Applications	10
4.4. Starting the Local AppFabric	11
4.5. Deploying and Running Applications using the AppFabric UI	11
4.6. Push-To-Cloud	12
5. AppFabric Programming Guide	13
5.1. AppFabric Core APIs.....	13
5.2. The Flow System	19
Sequential and Asynchronous Flowlet Execution.....	19
Flows and Instances.....	20

Getting Data In	20
5.3. The Transaction System	21
5.4. The Dataset System	22
Types of Datasets	22
Core Datasets - Tables	22
System Datasets	25
Custom Datasets.....	25
5.5. End-to-End Programming Example	27
Defining The Application	27
Defining The Flow	28
Implementing Flowlets.....	28
Implementing Custom Datasets	30
Implementing a Procedure.....	32
5.6. Testing Your Application	34
6. API and Tool Reference.....	36
6.1. Java APIs.....	36
6.2. REST APIs.....	36
REST Endpoint Port Configuration.....	36
Stream REST API	37
Data REST API	39
Procedure REST API	42
Monitor REST API	42
6.3. Command Line Tools.....	43
AppFabric	43
Data Client.....	43
Stream Client.....	44
AppFabric Client	45
6.4. Eclipse IDE Plugin	46
Prerequisites.....	46
Getting Eclipse and Setting up Plugin.....	46
Creating a Simple Application	46
Creating a Flow	48
Configure the Application.....	51
Running and Debugging Application	51
7. Conclusion	56

1. INTRODUCTION

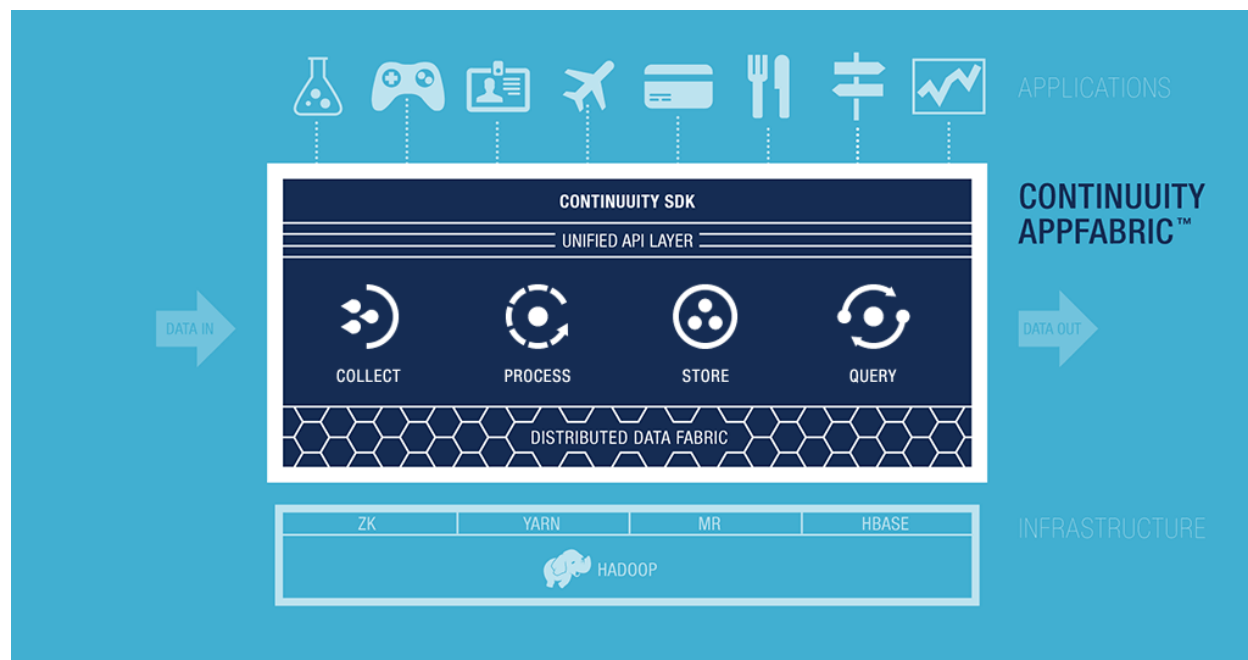
Developing, testing, running and scaling Big Data applications can be a difficult and complicated process requiring specialized expertise and large teams of engineers and operators. It is also a massive undertaking to educate yourself about the different projects support Big Data applications within the Hadoop ecosystem, understand how they fit together, and effectively decide which technologies best suit your specific use cases.

Continuity empowers you, the developer, by abstracting away unnecessary complexity and exposing the power of Big Data and Hadoop through higher-level abstractions, simple REST interfaces, powerful developer tools, and the Continuity AppFabric, a scalable and integrated runtime environment and data platform with a rich visual user interface. Using Continuity, developers can easily and quickly build, run, and scale their own Big Data applications, from prototype to production.

This guide is intended for developers and explains the major concepts and key capabilities supported by the Continuity AppFabric, including an overview of the core APIs, libraries, and the web UI. The Getting Started section will have you running your own instance of the AppFabric and deploying a sample application in minutes. To get you developing your own applications, the programming guide will deep-dive into the Core APIs and walk you through the implementation of an entire application, giving you an understanding of how the Continuity AppFabric capabilities enable you to quickly and easily build your own custom applications.

1.1. WHAT IS THE CONTINUUNITY APPFABRIC?

The Continuity AppFabric is a Java-based, integrated data and application framework that layers on top of Apache Hadoop, Apache HBase, and other components within the Hadoop ecosystem. It surfaces capabilities of the underlying infrastructure through simple Java and REST APIs and shields you from unnecessary complexity. Rather than piecing together different open source frameworks and runtimes to assemble your own Big Data infrastructure stack, Continuity provides an integrated platform, the AppFabric, that makes it easy to compose the different elements of your Big Data application: collecting, processing, storing, and querying data.



The production Continuity AppFabric is a distributed cloud platform, available as a hosted Virtual Private Cloud or in an On-Premise environment. For development, you can first run a local version of the AppFabric on your own

machine, which makes testing and debugging easy, and then you can self-provision a free, hosted Developer Sandbox to experience “Push-to-Cloud” functionality. Regardless of what version you use, your application code and your interactions with the AppFabric remain the same.

1.2. WHAT IS THE DEVELOPER SUITE?

The Continuity Developer Suite gives you everything you need to develop, test, debug and run your own Big Data applications: a complete set of APIs, libraries, and documentation, an IDE plugin, sample applications, and the local version of the AppFabric. The Developer Suite is your on-ramp to the distributed Continuity AppFabric, enabling you to develop locally and then push to a distributed AppFabric with a single click. Your interactions with the distributed AppFabric are the same as when you developed it the local AppFabric, but you can now control the scale of your application to meet its demands, in real-time, with no application downtime.

WHAT’S IN THE BOX?

The Continuity Developer Suite includes the AppFabric Software Development Kit (SDK) and the local version of the Continuity AppFabric.

APPFABRIC SDK

The SDK includes this developer guide, all of the Continuity APIs and libraries, Javadoc, command-line tools, Eclipse IDE plugin, and sample applications. See the Getting Started section for more details.

LOCAL APPFABRIC

The local version of the AppFabric is a fully functional but scaled-down single-node runtime environment that emulates the typically distributed and large-scale infrastructure (Hadoop and HBase) in a lightweight way. You run the local AppFabric on your own development machine, deploy your applications to it, and use the local UI to control and monitor it. You have direct access to your running application, making it easy to attach a debugger or profiler.

1.3. HOW TO BUILD APPS USING CONTINUITY

You build the core of your application in your own IDE using the Continuity Core Java APIs and libraries included in the AppFabric SDK. We help to get you started with an Eclipse plugin and sample projects, as well as a set of example applications that utilize the various features of the AppFabric.

Once the first version of your application has been built, you can deploy it to your local AppFabric using the local AppFabric UI or the Eclipse Plugin. Then you can begin the process of testing, debugging, and iterating on your application.

Getting data in and out of your application can be done programmatically using REST APIs or through the UI and the command line tools.

To test your application in a cloud environment you can deploy it to the Continuity Developer Sandbox.

When ready for production, your application can be easily deployed from your local machine to a distributed instance with no code changes or manual configuration. There it will be highly available and can be scaled to meet the dynamic demands of your app.

2. HELLO WORLD

Before going into the details of what the Continuity AppFabric is and how it works, here is simple code example for the curious developer, a “Hello World!” application. It produces friendly greetings, using one stream, one dataset, one flow and one procedure – the next section will introduce these concepts more formally, and Section 4 will explain all the APIs used here. The application:

- Receives names as real-time events on a stream;
- Processes the stream with a flow that stores each name in a key/value table;
- On request, reads the latest name from the key/value table and returns “Hello <name>!”

```
public class HelloWorld implements Application {
    @Override
    public ApplicationSpecification configure() {
        return ApplicationSpecification.Builder.with().
            setName("HelloWorld").
            setDescription("A Hello World program for the App Fabric").
            withStreams().add(new Stream("who")).
            withDataSets().add(new KeyValueTable("whom")).
            withFlows().add(new WhoFlow()).
            withProcedures().add(new Greeting()).
            build();
    }
    public static class WhoFlow implements Flow {
        @Override
        public FlowSpecification configure() {
            return FlowSpecification.Builder.with().
                setName("WhoFlow").
                setDescription("A flow that collects names").
                withFlowlets().add("saver", new NameSaver()).
                connect().fromStream("who").to("saver").
                build();
        }
    }
    public static class NameSaver extends AbstractFlowlet {
        @UseDataSet("whom") KeyValueTable whom;
        public void processInput(StreamEvent event) throws OperationException {
            byte[] name = Bytes.toBytes(event.getBody());
            if (name != null && name.length > 0) {
                whom.write(NAME, name);
            }
        }
    }
    public static class Greeting extends AbstractProcedure {
        @UseDataSet("whom") KeyValueTable whom;
        @Handle("greet")
        public void greet(ProcedureRequest req, ProcedureResponder responder) throws Exception {
            byte[] name = whom.read(NAME);
            String toGreet = name != null ? new String(name) : "World";
            responder.sendJson(new ProcedureResponse(SUCCESS, "Hello " + toGreet + "!"));
        }
    }
}
```

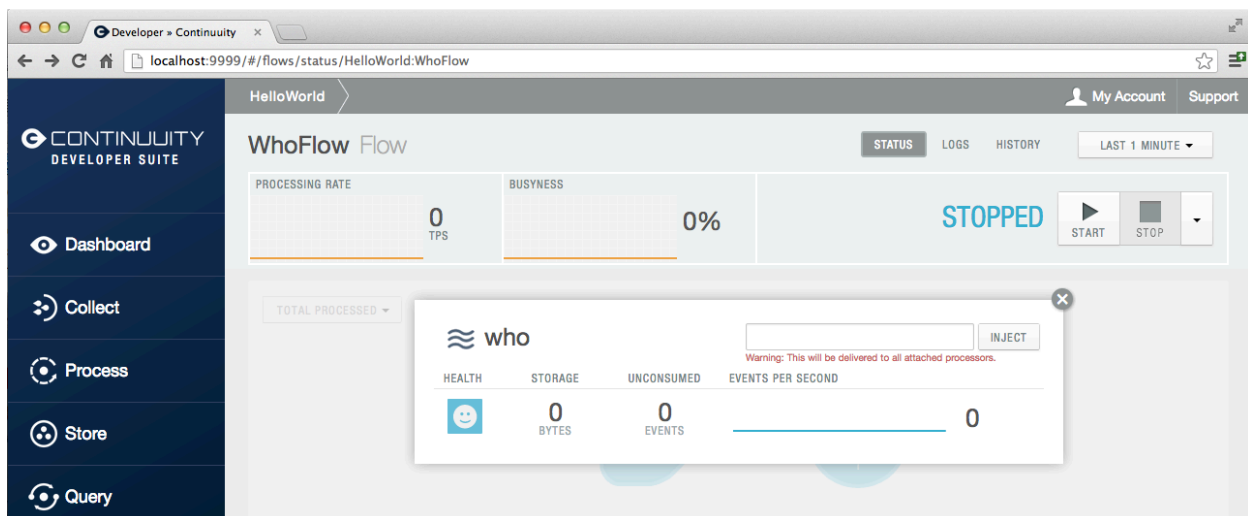
This code is included along with other examples in the Developer Suite. To see this application working, first build it from the examples directory.

```
> cd continuity-developer-suite-1.4.0/examples/HelloWorld
> ant
```

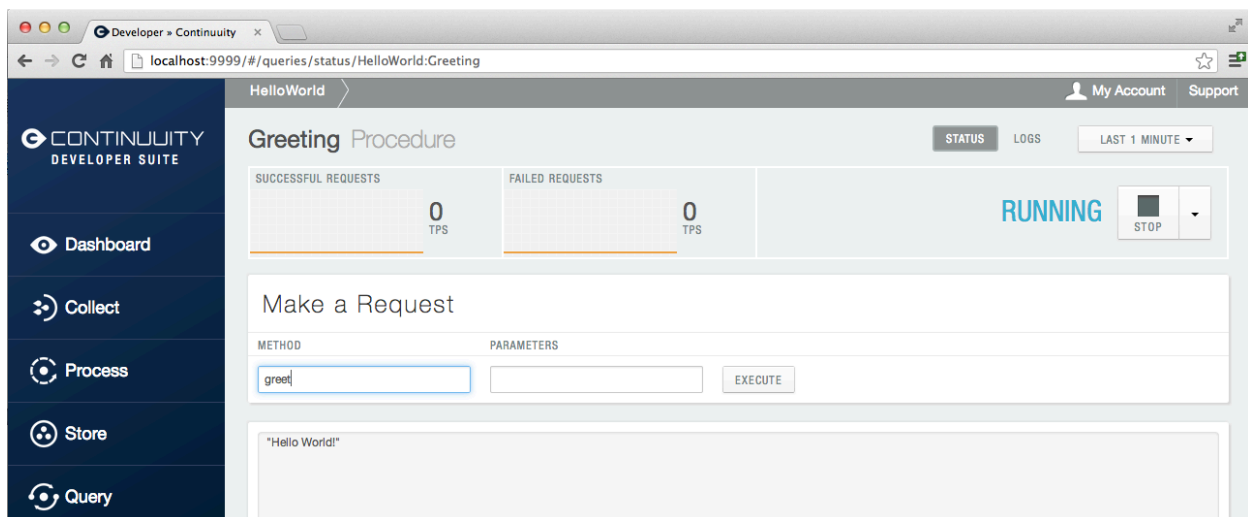
This creates archive named `HelloWorld.jar` in the same directory. To deploy the application, start the AppFabric:

```
> continuity-app-fabric start
```

Go to the local user interface at <http://localhost:9999/>, click on ADD AN APP on the right hand side, and drag the `HelloWorld.jar` onto the drop area. You will then see a dashboard showing one application named HelloWorld. Click on it to see the stream, flow, dataset and procedure that belong to this application. To send a name to the stream, click on the flow named WhoFlow and you will see a graphic rendering of the flow. Click the START button to start the flow, then click on the stream item labeled “who” and enter a name:

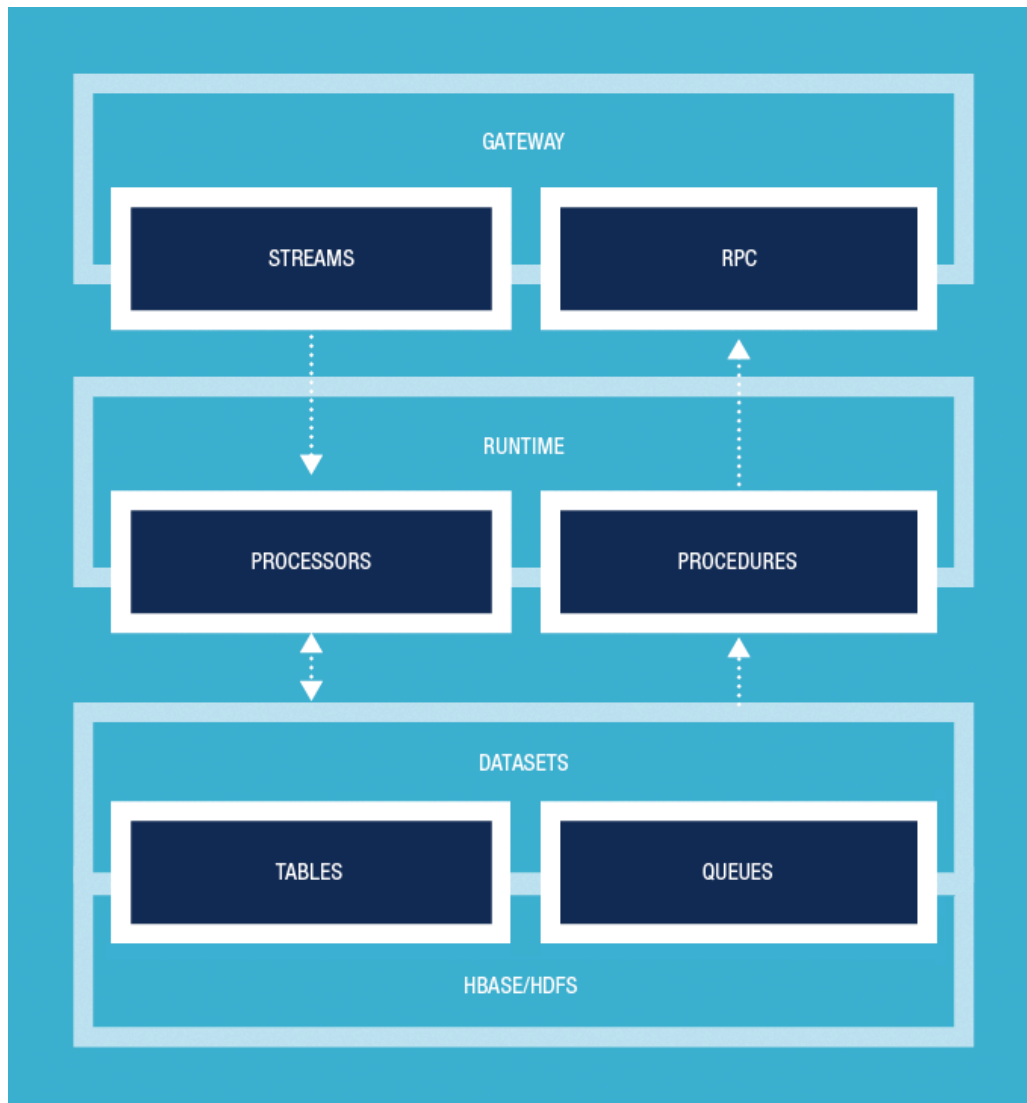


Now click on the Query icon on the left hand, and you see one procedure named Greeting. Click on it to get to the Procedure screen and click the START button to run the procedure. Now you can enter a query: Type `greet` into the METHOD box, and you will see the response:



3. UNDERSTANDING THE CONTINUUNITY APPFABRIC

The Continuity AppFabric is a unified Big Data application platform that brings various Big Data capabilities into a single environment and provides an elastic application runtime for user code. Data can be stored in both structured and unstructured forms; ingestion, processing, and serving can be done in real-time. Everything is elastically scalable.



The AppFabric provides **Streams** for simple data ingestion from any external system, **Processors** for performing elastically scalable real-time stream processing, **Datasets** for storing data in a simple and scalable way without worrying about formats and schema, and **Procedures** for exposing data to external systems as simple or complex interactive queries. These are grouped into **Applications** for configuring and packaging into deployable AppFabric artifacts.

As a developer, you will build Applications in Java using the Continuity Core APIs. Once your Application is deployed and running, you can easily interact with it from virtually any external system by accessing the streams, datasets, and procedures using REST or other network protocols.

3.1. COLLECT WITH STREAMS

Streams are the primary means for sending data from external systems into the AppFabric. You can write to streams easily using REST (see Section 6.2) or command-line tools (see Section 6.3), either one operation at a time or in batches. Each individual signal sent to a stream is stored as an **Event**, which is comprised of a body (blob of arbitrary binary data) and headers (map of strings for metadata).

Streams are identified by a *Unique Stream ID* string and must be explicitly created before being used. They can be created using a command-line tool (see Section 6.3), the Management UI, or programmatically within your Application (see Section 5.1). Data written to a stream can then be consumed by flows and processed in real-time as described below.

3.2. PROCESS WITH FLOWS

Flows are user-implemented real-time stream processors. They are comprised of one or more **Flowlets** that are wired together into a DAG (Directed Acyclic Graph). Flowlets pass **Data Objects** between one another; each flowlet is able to perform custom logic and execute data operations for each individual data object processed. All data operations happen in a consistent and durable way (more about this in Sections 5.2 and 0).

Flows are deployed into the AppFabric and hosted within containers, such that each instance of a flowlet runs in its own container. Each flowlet in the DAG can have multiple concurrent instances, each consuming a partition of the flowlet's inputs.

To get data into your flow, you can either connect the input of the flow to a stream, or you can implement a **Generator Flowlet**, which executes custom code in order to generate data or pull it from an external source.

To learn more about flows see Section 5.2.

3.3. STORE WITH DATASETS

Datasets are your interface to the AppFabric's storage engine, the DataFabric. Instead of requiring to manipulate data with low-level DataFabric APIs, datasets provide higher level abstractions and generic, reusable Java implementations of common data patterns.

TYPES OF DATASETS

The **core** dataset of the DataFabric is a **Table**. Other than in relational database systems, tables are not organized into rows with a fixed schema, and they are optimized for accessing and manipulating data at the column level. Tables allow for efficient storage of semi-structured data, data with unknown or variable schema, and sparse data.

All other datasets are built on top of the core datasets, that is, tables. For example, a dataset can implement a specific semantics around a table, such as key/value table, or a counter table. A dataset can also combine multiple tables into a more complex data pattern. For example, an indexed table can be implemented using one table for the data to index and a second table for the index.

In Section 0 you will learn how to implement your own, specific data patterns as **custom** datasets on top of tables. Because a number of useful datasets, including key/value tables, indexed tables and time series, are already included with the AppFabric, we call them **system** datasets.

To learn more about datasets see Section 0.

3.4. QUERY WITH PROCEDURES

Procedures allow you to make synchronous calls into the AppFabric from external systems and perform server-side processing on-demand, similar to a stored procedure in a traditional database. A procedure implements and

exposes a very simple API: method name (string) and arguments (map of strings). This implementation is then bound to a REST endpoint and can be called from any external system.

Procedures are typically used to post-process data at query time. This post-processing can include filtering, aggregations, or joins over multiple datasets – in fact, a procedure can perform all the same operations as a flowlet with the same consistency and durability guarantees. They are deployed into the same pool of application containers as flows, and you can run multiple instances to increase the throughput of requests.

We will learn more about procedures in section 5.1.

3.5. PACKAGE WITH APPLICATIONS

Applications are the highest-level concept and serve to specify and package all the components and configurations of your Big Data application. Within the application, you can explicitly indicate (and if necessary, create) your streams and datasets and declare all of the flows and procedures that make up the application.

3.6. APPFABRIC RUNTIME EDITIONS

The Continuuity AppFabric can be run in different modes: in-memory mode for unit testing, local mode for local testing, and distributed mode for staging and production. In addition, you have the option to get a free Developer Sandbox in the cloud. Regardless of the runtime edition, the AppFabric is fully functional and the code you develop never changes, however performance and scale are limited when using in-memory or local mode or a developer sandbox.

IN-MEMORY APPFABRIC

The in-memory AppFabric allows you to easily run the AppFabric for use in JUnit tests. In this mode, the underlying Big Data infrastructure is emulated using in-memory data structures and there is no persistence. There is no UI available for this mode.

LOCAL APPFABRIC

The local AppFabric allows you to run the entire AppFabric stack in a single JVM on your local machine and also includes a local version of the UI. The underlying Big Data infrastructure is emulated on top of your local file system and a relational database. All data is persisted.

See section 0 in the Getting Started section for more information on how to start and manage your local AppFabric.

DEVELOPER SANDBOX

The developer sandbox is a free version of the AppFabric that is hosted and operated in the Cloud. However, it does not provide the same scalability and performance as the fully distributed runtime. The developer sandbox is a good way to experience all the features of the “push-to-cloud” functionality and a hosted AppFabric without paying for the full, distributed edition.

DISTRIBUTED APPFABRIC

All other hosted versions of the Continuuity AppFabric, that is, the Virtual Private Cloud edition and the On-Premise edition, run in a fully distributed mode. This includes distributed and highly available deployments of the underlying Hadoop infrastructure as well as the other system components of the AppFabric. Production applications should always be run on a distributed AppFabric.

To self-provision your own Developer Sandbox simply go to your Account Home page located at <https://accounts.continuuity.com>. To learn more about getting your own production distributed AppFabric, check out <http://www.continuuity.com/products>.

4. GETTING STARTED WITH THE LOCAL APPFABRIC

Before diving into the hands-on development of your application, it might be useful to build, deploy, and run one of the provided sample applications using the local AppFabric and its UI.

4.1. PREREQUISITES

The only prerequisite for running the local AppFabric is that you need to have Node™ installed. You can get the latest version at <http://nodejs.org>.

4.2. UNPACKING THE DEVELOPER SUITE

The Continuity Developer Suite is bundled as a zip file and contains everything you need to build and run Big Data applications on your local machine. It includes the AppFabric API Jar, documentation, tools, IDE plugin, and sample applications:

README	(a file you should read)
LICENSE	(the Developer Suite license)
continuity-api-1.4.0.jar	(the API jar for AppFabric)
bin/continuity-app-fabric	(Single-Node AppFabric Daemon)
bin/data-format	(Single-Node Data Format Tool)
bin/stream-client	(Command-Line Stream Client)
bin/data-client	(Command-Line Dataset Client)
bin/app-fabric-client	(Command-Line App Fabric Client)
docs/api	(API JavaDocs)
conf/logback.xml	(Local AppFabric Log Configuration)
conf/continuity-site.xml	(Local AppFabric Configuration)
web-app	(AppFabric UI application)
examples/WordCount	(Sample Word Count Application)
examples/CountAndFilterWords	(Sample Word Filter Application)
examples/CountCounts	(Sample Number Counter Application)
examples/CountOddAndEven	(Sample Odd/Even Number Application)
examples/CountRandom	(Sample Random Number Generator App)
examples/CountTokens	(Sample String Counting Application)
examples/HelloWorld	(Sample Hello World Application)
examples/SimpleWriteAndRead	(Sample Dataset using Application)
examples/build.xml	(Ant build.xml for building examples)

4.3. BUILDING EXAMPLE APPLICATIONS

Building the example applications is simple using ant:

```
> cd ~/continuity-developer-edition-1.4.0/examples
> ant
```

This will generate a JAR file for each of the sample apps. You can also individually build a single example:

```
> cd ~/continuity-developer-edition-1.4.0/examples/WordCount
> ant
```

4.4. STARTING THE LOCAL APPFABRIC

To start the local AppFabric, simply run the `continuuity-app-fabric` daemon with the start command:

```
> cd ~/continuuity-developer-edition-1.4.0/  
> ./bin/continuuity-app-fabric start
```

Your local AppFabric is now up-and-running. You can check on its status or stop it:

```
> ./bin/continuuity-app-fabric status  
> ./bin/continuuity-app-fabric stop
```

The URL for the local AppFabric UI is displayed on your screen. It is also accessible via <http://localhost:9999>.

4.5. DEPLOYING AND RUNNING APPLICATIONS USING THE APPFABRIC UI

Now that the local AppFabric instance is running and you have accessed your local (but empty) AppFabric UI, you can easily deploy and run one of the bundled sample apps. In this example, deploy the `WordCount` application by drag-and-dropping the `WordCount.jar` file onto the UI.

1. Click on “Create Application” to open the New Application dialog box
2. Drag-and-drop your `WordCount.jar` onto the New Application dialog
3. Your application is now uploaded, deployed, and verified

With your Application deployed, you can now start its flow and the procedure.

1. Click on the `WordCount` application in the Dashboard Application list
2. You should see all of the components of this application: a stream, a flow, four datasets, and a procedure.
3. Click on the `WordCounter` flow to open the flow visualization page
4. Press the START button in the top-right corner to run the flow

The flow is running, so now it can begin to process incoming data from the stream. In order to send a sample event using the UI:

1. Click on the `wordStream` stream icon to open the stream dialog.
2. Type a string of words into the top-right textbox and click the INJECT button
3. Watch the event get processed by the flowlets in the DAG

You can also send events to a stream using REST (see Section 6.2) or the command-line (see Section 6.3).

After you have processed some data from the stream and with the flow, use the procedure to read the results of the processing:

1. Return to the `WordCount` application page
2. Click on the `RetrieveCount` procedure to open the Procedure page
3. Press the START button in the top-right corner to run the procedure

The procedure is now running and ready to accept new requests. Procedures bind to REST interfaces so it is easy to query them using any HTTP-based tools or libraries (see Section 6.2). Additionally, you can use the procedure UI to send REST requests and receive the response directly in the UI.

1. Enter `getCount` in the METHOD box
2. Enter the JSON string `{"word": "..."}` in the PARAMETERS box (use a word that you sent to the stream)
3. Press the EXECUTE button
4. JSON results will be displayed in the large text box

4.6. PUSH-TO-CLOUD

After you have developed and tested your application in the local AppFabric, you can promote it to a remote instance of the AppFabric in the cloud. Currently, that instance is only a Developer Sandbox. With the sandbox you can experience the push-to-cloud feature without deploying a full, distributed AppFabric. You can create your developer sandbox at <https://accounts.continuity.com>. At that time you will also receive an **API Token** that authenticates you with the sandbox. You need this API token to promote an application to the cloud.

1. Return to the [WordCount](#) application page.
2. Click on the PUSH button in the top-right corner to open the Push-to-Cloud dialog.
3. Enter your API key. Your sandbox should automatically appear in the dialog.
4. Click on the PUSH button.

Your application will now be promoted to your sandbox. Then you can use the UI of the sandbox to interact with your application in the same way as you did with the local UI.

5. APPFABRIC PROGRAMMING GUIDE

This section dives into more detail around each of the different AppFabric core capabilities - Streams, Flows, Datasets, and Procedures - and how you work with them in Java to build your Big Data Application.

First there is an overview of all of the high-level concepts and core Java APIs. Then a deep-dive into the Flow System, Procedure System, Datasets, and Transactions will give an understanding of how these systems function. Finally an example application will be implemented to help illustrate these concepts and describe how an entire application is built.

5.1. APPFABRIC CORE APIs

Application An application is a collection of streams, flows, datasets, and procedures. To create an application you implement the `Application` interface and its `configure()` method. This allows you to specify the application metadata, declare and configure the streams and datasets used, and add the associated flows and procedures.

```
public class MyApplication implements Application {
    @Override
    public ApplicationSpecification configure() {
        return ApplicationSpecification.Builder.with()
            .setName("myApp")
            .setDescription("my sample application")
            .withStreams().add(...) ...
            .withDataSets().add(...) ...
            .withFlows().add(...) ...
            .withProcedures().add(...) ...
            .build();
    }
}
```

You can also specify that an application does not use a stream:

```
.setDescription("my sample application")
.noStream().
.withDataSets().add(...) ...
```

and similarly for all of the other constructs.

Stream Streams are the primary means for pushing data into the AppFabric. You can specify a stream in your application as follows:

```
.withStreams().add(new Stream("myStream")) ...
```

Flow Flows are a collection of connected flowlets, wired into a DAG. To create a flow you implement the `Flow` interface and its `configure()` method. This allows you to specify the flow's metadata, flowlets, flowlet connections, stream to flowlet connections, and any datasets used in the flow using a `FlowSpecification`:

```
class MyExampleFlow implements Flow {
    @Override
    public FlowSpecification configure() {
        return FlowSpecification.Builder.with()
            .setName("mySampleFlow")
            .setDescription("Flow for showing examples")
            .withFlowlets()
            .add("flowlet1", new MyExampleFlowlet())
            .add("flowlet2", new MyExampleFlowlet2())
    }
}
```

```

        .connect()
        .fromStream("myStream").to("flowlet1")
        .from("flowlet1").to("flowlet2")
        .build();
    }

```

Flowlet

The basic building blocks of a flow, flowlets represent each individual processing node within a flow. Flowlets consume data objects from their inputs and execute custom logic on each data object, allowing you to perform data operations as well as emit data objects to the flowlet's outputs. Flowlets also specify an `initialize()` method, which is executed at the startup of each instance of a flowlet before it receives any data.

The example below shows a flowlet that reads Double values, rounds them and emits the results. It has a very simple configuration method, and does nothing for initialization and destruction (see below for a simpler way to declare these methods):

```

class RoundingFlowlet implements Flowlet {

    @Override
    public FlowletSpecification configure() {
        return FlowletSpecification.Builder.with().
            setName("round").
            setDescription("a rounding flowlet").
            build();
    }

    @Override
    public void initialize(FlowletContext context) throws FlowletException {
    }

    @Override
    public void destroy() {
    }
}

```

The most interesting method of this flowlet is `round()`. It is the method that does the actual processing. It uses an output emitter to send data to its output. This is the only way that a flowlet can emit output:

```

    OutputEmitter<Long> output;

    @ProcessInput
    public void round(Double number) {
        output.emit(Math.round(number));
    }
}

```

Note that the flowlet declares the `OutputEmitter` but does not initialize it: The flow system injects its implementation at runtime. Note also that the method is annotated with `@ProcessInput` – this tells the flow system that this method can process input data. Another way to define this method is to have its name start with `process` – in which case no annotation is needed:

```

    public void processDouble(Double number) {
        output.emit(Math.round(number));
    }
}

```

You can also overload the process method of a flowlet by adding multiple methods with different input types. When an input object comes in, the flowlet will call the method that matches the object's type.

```
OutputEmitter<Long> output;

@ProcessInput
public void round(Double number) {
    output.emit(Math.round(number));
}
@ProcessInput
public void round(Float number) {
    output.emit((long) Math.round(number));
}
```

If you define multiple process methods, a method can be selected based on the input object's origin, that is, the name of a stream or the name of an output of a flowlet. A flowlet that emits data can specify this name using an annotation on the output emitter (in the absence of this annotation, the name of the output defaults to "out"):

```
@Output("code")
OutputEmitter<String> out;
```

Data objects emitted through this output can then be directed to a process method by annotating the method with the origin name:

```
@ProcessInput("code")
public void tokenizeCode(String text) {
    ... // perform fancy code tokenization
}
```

A process method can have an additional parameter, the input context. The input context provides information about the input object, such as its origin and the number of times the object has been retried (see Section 5.2 for the retry logic of a flowlet). For example, the following flowlet tokenizes text in a smart way and it uses the input context to decide what tokenizer to use.

```
@ProcessInput
public void tokenize(String text, InputContext context) throws Exception {
    Tokenizer tokenizer;
    // if this failed before, fall back to simple white space
    if (context.getRetryCount() > 0) {
        tokenizer = new WhiteSpaceTokenizer();
    }
    // is this code? If it's origin is named "code", then assume yes
    else if ("code".equals(context.getOrigin())) {
        tokenizer = new CodeTokenizer();
    }
    else {
        // use the smarter tokenizer
        tokenizer = new NaturalLanguageTokenizer();
    }
    for (String token : tokenizer.tokenize(text)) {
        output.emit(token);
    }
}
```

Type Projection Flowlets perform an implicit projection on the input objects if they do not match exactly what the process method accepts as arguments. This allows you to write a single process method that can accept multiple **compatible** types. For example, if you have a process method:

```
@ProcessInput
count(String word) {
    ...
}
```

and you send data of type `long` to this flowlet, then that type does not exactly match what the process method expects. You could now write another process method for long numbers:

```
@ProcessInput
count(Long number) {
    count(number.toString());
}
```

and you could do that for every type that you might possibly want to count, but that would be rather tedious. Type projection does this for you automatically: If no process method is found that matches the type of an object exactly, it picks a method that is compatible with the object. In this case, because `long` can be converted into a `String`, it is compatible with the original process method. Other compatibilities include:

- Every primitive type that can be converted to a string is compatible with `String`.
- Any numeric type is compatible with numeric types that can represent it. For example, `int` is compatible with `long`, `float` and `double`, and `long` is compatible with `float` and `double`, but `long` is not compatible with `int` because `int` cannot represent every long value.
- A byte array is compatible with a `ByteBuffer` and vice versa.
- A collection of type A is compatible with a collection of type B, if A is compatible with B. Here, a collection can be an array or any Java `Collection`. Hence, a `List<Integer>` is compatible with a `String[]` array.
- Two maps are compatible if their underlying types are compatible. For example, a `TreeMap<Integer, Boolean>` is compatible with a `HashMap<String, String>`.
- Other Java objects can be compatible if their fields are compatible. For example, in the following class `Point` is compatible with `Coordinate`, because all common fields between the two classes are compatible. When projecting from `Point` to `Coordinate`, the `color` field is dropped, whereas the projection from `Coordinate` to `Point` will leave the `color` field as `null`.

```
class Point {
    private int x;
    private int y;
    private String color;
}

class Coordinates {
    int x;
    int y;
}
```


Type projections help you keep your code generic and reusable. They also interact well with inheritance: If a flowlet can process a specific object class, then it can also process any subclass of that class.

Stream Event A stream event is a special type of object that comes in via streams. It consists of a set of headers represented by a map from string to string, and a byte array as the body of the event. To consume a stream with a flow, define a flowlet that processes data of type `StreamEvent`:

```
class StreamReader extends AbstractFlowlet {  
    ...  
    public void processEvent(StreamEvent event) {  
        ...  
    }  
}
```

Generator A special case of a flowlet is a generator flowlet. The difference from a standard flowlet is that a generator has no inputs. Instead of a process method, it defines a `generate()` method to emit data. This can be used, for instance, to generate test data, or connect to an external data source and pull data from there.

For example, the following generator flowlet emits random numbers. It extends the `AbstractFlowlet` class that provides simple default implementations of the `configure`, `initialize` and `destroy` methods:

```
class RandomGenerator extends AbstractFlowlet implements GeneratorFlowlet {  
  
    private OutputEmitter<Double> output;  
    private Random random = new Random();  
  
    @Override  
    public void generate() {  
        this.output.emit(random.nextDouble());  
    }  
}
```

Because this generator flowlet has an output of type `Double`, it can be connected to a `DoubleRoundingFlowlet`, which has a process method that accepts `Double`.

Connection There are multiple ways to connect the flowlets of a flow. The most common form is to use the flowlet name. Because the name of each flowlet defaults to its class name, you can do the following when building the flow specification:

```
.withFlowlets()  
    .add(new RandomGenerator())  
    .add(new RoundingFlowlet())  
.connect()  
    .fromStream("RandomGenerator").to("RoundingFlowlet")
```

If you have two flowlets of the same class you can give them explicit names:

```
.withFlowlets()  
    .add("random", new RandomGenerator())  
    .add("generator", new RandomGenerator())  
    .add("rounding", new RoundingFlowlet())  
.connect()  
    .fromStream("random").to("rounding")
```

Procedure

Procedures are a way to make calls into the AppFabric from external systems and perform arbitrary server-side processing on-demand.

To create a procedure you implement the `Procedure` interface, or more conveniently, you extend the `AbstractProcedure` class. It is configured and initialized similarly to a flowlet but instead of a process method, you define a handler method that, given a method name and map of string arguments, translated from an external request, sends a response. The most generic way to send a response is to obtain a `Writer` and stream out the response as bytes. You should make sure to close the writer when you are done:

```
class HelloWorld extends AbstractProcedure {

    @Handle("hello")
    public void wave(ProcedureRequest request,
                    ProcedureResponder responder) throws IOException {
        String hello = "Hello " + request.getArgument("who");
        ProcedureResponse.Writer writer =
            responder.stream(new ProcedureResponse(SUCCESS));
        writer.write(ByteBuffer.wrap(hello.getBytes())).close();
    }
}
```

This uses the most generic way to create the response, which allows you to send arbitrary byte content as the response body. In many case, you will actually respond in JSON. Procedures have a convenience methods for this:

```
// return a JSON map
Map<String, Object> results = new TreeMap<String, Object>();
results.put("totalWords", totalWords);
results.put("uniqueWords", uniqueWords);
results.put("averageLength", averageLength);
responder.sendJson(new ProcedureResponse(Code.SUCCESS), results);
```

There is also a convenience method to respond with an error message:

```
@Handle("getCount")
public void getCount(ProcedureRequest request, ProcedureResponder responder) {
    String word = request.getArgument("word");
    if (word == null) {
        responder.error(Code.CLIENT_ERROR,
            "Method 'getCount' requires argument 'word'");
        return;
    }
}
```

Dataset

Datasets are the way you store and retrieve data. If your application uses a dataset, then you must declare it in the application specification. For example, to specify that your application uses a `KeyValueTable` dataset named “myCounters”, you write:

```
public ApplicationSpecification configure() {
    return ApplicationSpecification.Builder.with()
        ...
        .withDataSets().add(new KeyValueTable("myCounters"))
        ...
}
```

In order to use the dataset inside a flowlet or a procedure, you rely on the runtime system to inject an instance of the dataset. You do that with an annotation:

```

Class myFlowlet extends AbstractFlowlet {

    @UseDataSet("myCounters")
    private KeyValueTable counters;
    ...
    void process(String key) throws OperationException {
        counters.increment(key.getBytes());
    }
}

```

The runtime system reads the dataset specification for the key/value table named “myCounters” from the metadata store and injects a functional instance of the dataset class into your code.

You can also implement your own datasets by extending the [DataSet](#) base class or extending any other existing type of dataset. More about this in the programming example in Section 0

Logging

The AppFabric supports logging from flows and procedures using standard SLF4J APIs. For instance, in a flowlet you can write:

```

private static Logger LOG = LoggerFactory.getLogger(WordCounter.class);
...
public void process(String line) {
    LOG.debug(this.getContext().getName() + ": Received line " + line);
    ... // processing
    LOG.debug(this.getContext().getName() + ": Emitting count " + wordCount);
    output.emit(wordCount);
}

```

The log messages emitted by your application code can be viewed in two different ways:

- All log messages of an application can be viewed in the UI, by clicking on the “Logs” button in the Flow and Procedure screens.
- In the local AppFabric, application log messages are also written to the system log files along with the messages emitted by the AppFabric itself. These files are not available for viewing in the developer sandbox.

5.2. THE FLOW SYSTEM

Flows are user-implemented real-time stream processors. They are comprised of one or more flowlets that are wired together into a DAG (Directed Acyclic Graph). Flowlets pass data between one another; each flowlet is able to perform custom logic and execute data operations for each individual data object it processes.

A flowlet processes the data objects from its input one by one. If a flowlet has multiple inputs, they are consumed in a round-robin fashion. When processing a single input object, all operations, including the removal of the object from the input, and emission of data to the outputs, are executed in a transaction. This provides us with ACID properties, and helps assure a unique and core property of the flow system: It guarantees atomic and exactly-once-processing of each input object by each flowlet in the DAG. See Section 0 to learn more about transactions.

SEQUENTIAL AND ASYNCHRONOUS FLOWLET EXECUTION

A flowlet processes the data from its inputs one object at a time, by repeating the following steps:

1. An object is dequeued from one of the inputs. This does not completely remove it from the input, but marks it as in-progress.
2. The matching [process\(\)](#) method is selected and invoked, in a new transaction. The process method can perform dataset operations and emit data to its outputs.

3. Once the `process()` method returns, the transaction is committed, and:
 - a. If the transaction **fails**, all the dataset operations are rolled back, and all emitted output data objects is discarded. The `onFailure()` callback of the flowlet is invoked, which allows you to retry or ignore
 - i. If you retry, `process()` will be invoked with the same input object again
 - ii. If you ignore, the failure will be ignored and the input object is permanently removed from the input.
 - b. If the transaction **succeeds**, all dataset operations are persistently committed, all emitted output data is sent downstream, and the input object is permanently removed from the input. Then the `onSuccess()` callback of the flowlet is invoked

By default, steps 1, 2 and 3 above happen in sequence, one input object at a time. You can increase the throughput of the flowlet by declaring it as asynchronous:

```
@Async
class MyFlowlet implements Flowlet {
    ...
}
```

Then the three steps happen concurrently: While the transaction of one data object is committing, the flowlet already processes the next object from the input, and yet another object is already being read from the input at the same time. However, you should be aware that processing now happens in overlapping transactions, and the probability of write conflicts increases – especially if the duration of the transactions is long (The transaction system has special ways to avoid conflicts on the flowlet’s inputs). Read more about Transactions in section 0.

FLows AND INSTANCES

You can have one or *more* instances of any given flowlet, each consuming a disjoint partition of each input. You can control the number of instances programmatically, using the UI, or using the command line interface. This enables you to shape your application to meet capacity at runtime the same way you will do in production. In the single-node version provided with the Developer Suite, multiples instances of a flowlet are run in threads, so in some cases actual performance may not be affected. In production, each instance is a separate JVM with independent compute resources.

GETTING DATA IN

Input data can be pushed to a flow using streams or pulled from within a flow using a generator flowlet.

- A **Generator Flowlet** actively retrieves or generates data, and the logic to do so is coded into the flowlet’s `generate()` method. For instance, a generator flowlet can connect to the Twitter fire hose or another external data source.
- A **Stream** is passively receiving events from outside (remember that streams exist outside the scope of a flow). To consume a stream, connect the stream to a flowlet that implements a process method for `StreamEvent`. This is useful when your events come from an external system that can push data using REST calls. It is also useful when you’re developing and testing your app, because your test driver can send mock data to the stream that covers all your test cases.

5.3. THE TRANSACTION SYSTEM

A flowlet processes the data objects from its inputs one at a time. While processing a single input object, all operations, including the removal of the data from the input, and emission of data to the outputs, are executed in a transaction. This provides us with ACID properties:

- The process method runs under read **isolation** to ensure that it does not see dirty writes (uncommitted writes from concurrent processing) in any of its reads. It does see, however, its own writes.
- A failed attempt to process an input object leaves the DataFabric in a **consistent** state, that is, it does not leave partial writes behind.
- All writes and emission of data are committed **atomically**, that is, either all of them or none of them are persisted.
- After processing completes successfully, all its writes are persisted in a **durable** way.

In case of failure, the state of the DataFabric is unchanged and therefore, processing of the input object can be reattempted. This ensures exactly-once processing of each object.

The AppFabric uses Optimistic Concurrency Control (OCC) to implement transactions. Unlike most relational databases that use locks to prevent conflicting operations between transactions, under OCC we allow these conflicting writes to happen. When the transaction is committed, we can detect whether it has any conflicts: namely if during the lifetime of this transaction, another transaction committed a write for one the same keys that this transaction has written. In that case, the transaction is aborted and all of its writes are rolled back.

In other words: If two overlapping transactions modify the same row, then the transaction that commits first will succeed, but the transaction that commits last is rolled back due to a write conflict.

Optimistic Concurrency Control is lockless and therefore avoids problems such as idle processes waiting for locks, or even worse, deadlocks. However, it comes at the cost of rollback in case of write conflicts. We can only achieve high throughput with OCC if the number of conflicts is small. It is therefore a good practice to reduce the probability of conflicts where possible:

- Keep transactions short. The AppFabric attempts to delay the beginning of each transaction as long as possible. For instance, if your flowlet only performs write operations, but no read operations, then all writes are deferred until the process method returns. They are then performed and transacted, together with the removal of the processed object from the input, in a single batch execution. This minimizes the duration of the transaction.
- However, if your flowlet performs a read, then the transaction must begin at the time of the read. If your flowlet performs long-running computations after that read, then the transaction runs longer, too, and the risk of conflicts increases. It is therefore a good practice to perform reads as late in the process method as possible.
- There are two ways to perform an increment: As a write operation that returns nothing, or as a read-write operation that returns the incremented value. If you perform the read-write operation, then that forces the transaction to begin, and the chance of conflict increases. Unless you depend on that return value, you should always perform an increment as a write operation.

Keeping these guidelines in mind will help you write more efficient code.

5.4. THE DATASET SYSTEM

Datasets are your interface to the DataFabric. Instead of requiring to manipulate data with low-level DataFabric APIs, datasets provide higher level abstractions and generic, reusable Java implementations of common data patterns. A dataset represents both the API and the actual data itself. In other words, a dataset class is a reusable, generic Java implementation of a common data pattern. A dataset Instance is a named collection of data with associated metadata, and it is manipulated through a Dataset Class.

TYPES OF DATASETS

A dataset is a Java class that extends the abstract `DataSet` class with its own, custom methods. The implementation of a dataset typically relies on one or more underlying (embedded) datasets. For example, the `IndexedTable` dataset can be implemented by two underlying `Table` datasets, one holding the data itself, and one holding the index. We distinguish three categories of datasets: core, system, and custom datasets:

- The **core** dataset of the DataFabric is a **Table**. Its implementation is hidden from developers and it may use private DataFabric interfaces that are not available to you.
- A **custom** dataset is implemented by you and can have arbitrary code and methods. It is typically built around one or more tables (or other datasets) to implement a more specific data pattern. In fact, a custom dataset can only interact with the DataFabric through its underlying datasets.
- A **system** dataset is bundled with the AppFabric but implemented in the same way as a custom dataset, relying on one or more underlying core or system datasets. The key difference between custom and system datasets is that system datasets are implemented (and tested) by Continuuity and as such they are reliable and trusted code.

Each dataset instance has exactly one dataset class to manipulate it - we think of the class as the type or the interface of the dataset. Every instance of a dataset has a unique name (unique within the account that it belongs to), and some metadata that defines its behavior. For example, every `IndexedTable` has a name and indexes a particular column of its primary table: The name of that column is a metadata property of each instance.

Every applications must declare all datasets that it uses in its application specification. The specification of the dataset must include its name and all its metadata, including the specifications of its underlying datasets. This allows creating the dataset - if it does not exist yet - and storing its metadata at the time of deployment of the application. Application code (for example, a flow or procedure) can then use a dataset by giving only its name and type - the runtime system can use the stored metadata to create an instance of the dataset class with all required metadata.

CORE DATASETS - TABLES

Tables are the only core dataset, and all other datasets are built using one or more underlying tables. A table is similar to tables in a relational database, but with a few key differences:

- Tables have no fixed schema. Unlike relational tables where every row has the same schema, every row of a table can have a different set of columns.
- Because the set of columns is not known ahead of time, the columns of a row do not have a rich type. All column values are byte arrays and it is up to the application to convert them to and from rich types. The column names and the row key are also byte arrays.
- When reading from a table, one need not know the names of columns: The read operation returns a map from column name to column value. It is, however, possible to specify exactly which columns to read.
- Tables are organized in a way that the columns of a row can be read and written independently of other columns, and columns are ordered in byte-lexicographic order. They are therefore also called Ordered Columnar Tables.

Interface

The table interface provides methods to perform read and write operations, plus a special increment method:

```
public class Table extends DataSet {

    public Table(String name);

    public OperationResult<Map<byte[], byte[]>> read(Read read)
        throws OperationException;

    public void write(WriteOperation op)
        throws OperationException;

    public Map<byte[], Long> increment(Increment increment)
        throws OperationException;
}
```

All operations can throw an `OperationException`. In case of success, the read operation returns an `OperationResult`, which is a wrapper class around the actual return type. In addition to carrying the result value it can indicate that no result was found and the reason why.

```
class OperationResult<ReturnType> {
    public boolean isEmpty();
    public String getMessage();
    public int getStatus();
    public ReturnType getValue();
}
```

Read

To read from a table you specify a row key and optionally the columns to read:

```
Table table;
// reads all columns
result = table.read(new Read(rowkey));
// reads only one column
result = table.read(new Read(rowkey, column1));
// reads specified set of columns
result = table.read(new Read(rowkey, new byte[][] { col1, col2, col3 }));
// reads all columns from colA up to but excluding colB
result = table.read(new Read(rowkey, colA, colB));
// reads upto 100 columns from colA up to but excluding colB
result = table.read(new Read(rowkey, colA, colB, 100));
// read all columns up to, but excluding colB
result = table.read(new Read(rowkey, null, colB));
// read upto 100 columns starting with colA
result = table.read(new Read(rowkey, colA, null, 100));
// read all columns in the row
result = table.read(new Read(rowkey, null, null));
```

Write

There are four types of write operations: Write, Delete, Swap and Increment. A write specifies a row key, a set of column keys and the same number of column values:

```
// write a new value to a single column
table.write(new Write(rowkey, column1, value1));
// write multiple columns
table.write(new Write(rowkey, new byte[][] { col1, col2 },
                    new byte[][] { val1, val2 }));
```

Note that the write does not replace the entire row: It only overwrites the values of the specified columns whereas existing values of other columns remain unmodified. If you want remove columns from a row, you must use a Delete operation (see below).

Increment

An Increment interprets each column as an 8-byte (long) integer and increments it by a given value. If the column does not exist, it is created with the value that was given as the increment. If the existing value of the column is not 8 bytes long, the operation will throw an exception.

```
// write a new value to a single column
table.write(new Increment(rowkey, countCol, 1L));
// write multiple columns
table.write(new Increment(rowkey, new byte[][] { col1, col2 },
                                     new long[] { 5, 10 }));
```

Note that this does not return the incremented values. If you want to use the result of the increment operation in subsequent code, you can use the `incrementAndGet` method. For example, to emit the incremented value to an output:

```
Map<byte[], Long> res =
    table.incrementAndGet(new Increment(rowkey, col, 1L));
Long incrementedValue = res.get(col);
output.emit(incrementedValue);
```

Delete

A delete removes the specified columns from a row. Note that this does not remove the entire row. If you want to delete an entire row, you need to know its columns and specify each one.

```
// delete a single column
table.write(new Delete(rowkey, column1));
// delete multiple columns
table.write(new Delete(rowkey, new byte[][] { col1, col2 }));
```

Swap

A swap operation compares the existing value of a column with an expected value, and if it matches, replaces it with a new value. This is useful to verify that a value has not changed since it was read. If it does not match, the operation fails and throws an exception. Read more about Optimistic Concurrency Control in section 0.

```
// read a user profile
result = table.read(new Read(userkey, profileCol));
oldProfile = result.getValue().get(profileCol);
...
newProfile = manipulate(oldProfile, ...);
// fails if somebody else has updated it in the mean time
table.swap(userkey, profileCol, oldProfile, newProfile);
```


SYSTEM DATASETS

The AppFabric comes with several system-defined datasets, including key/value tables, indexed tables and time series. Each of them is defined with the help one or more embedded tables, but defines its own interface. For example:

- The [KeyValueTable](#) implements a key/value store as a table with a single column.
- The [IndexedTable](#) implements a table with a secondary key using two embedded tables, one for the data and one for the secondary index.
- The [TimeseriesTable](#) uses a table to store keyed data over time and allows querying that data over ranges of time.

See the Java documentation of these classes to learn more about these datasets.

CUSTOM DATASETS

You can define your own dataset classes to implement common data patterns specific to your code. We will illustrate how to define your own dataset by means of an example. Suppose we want to define a counter table that in addition to counting words also counts how many unique words it has seen. The dataset will be built on top two underlying datasets, a [KeyValueTable](#) to count all the words and a core table for the unique count:

```
public class UniqueCountTable extends DataSet {  
  
    private Table uniqueCountTable;  
    private KeyValueTable wordCountTable;
```

In the constructor we take a name and create the two underlying datasets. Note that we use different names for the two tables, both derived from the name of the unique count table.

```
    public UniqueCountTable(String name) {  
        super(name);  
        this.uniqueCountTable = new Table("unique_count_" + name);  
        this.wordCountTable = new KeyValueTable("word_count_" + name);  
    }
```

Like most other components of an application, the dataset must implement a `configure()` method that returns a specification. In the specification we save metadata about the dataset (such as its name) and the specifications of the embedded datasets obtained by calling their respective `configure` methods.

```
@Override  
public DataSetSpecification configure() {  
    return new DataSetSpecification.Builder(this)  
        .dataset(this.uniqueCountTable.configure())  
        .dataset(this.wordCountTable.configure())  
        .create();  
}
```

So far, we have written all code needed to use the dataset in an application specification, that is, at application deploy time. At that time the dataset specification returned by `configure()` is stored with the application's metadata. At runtime the dataset must be available in all places that use the dataset, that is, in all instances of flowlets or procedures. To accomplish this the dataset is instantiated by reading its specification from the metadata store and calling the dataset's constructor that takes a dataset specification as argument. Every dataset class must implement such a constructor; otherwise, it will not be functional at runtime.

```

public UniqueCountTable(DataSetSpecification spec) {
    super(spec);
    this.uniqueCountTable = new Table(
        spec.getSpecificationFor("unique_count_" + this.getName()));
    this.wordCountTable = new KeyValueTable(
        spec.getSpecificationFor("word_count_" + this.getName()));
}

```

Now we can begin with the implementation of the dataset logic. We begin with a constants.

```

/** Row and column name used for storing the unique count */
private static final byte [] UNIQUE_COUNT = Bytes.toBytes("unique");

```

The dataset stores a counter for each word in its own row of the word count table, and for every word it increments its counter. If the resulting value is 1, then this was the first time we encountered the word, hence we have new unique word and we increment the unique counter.

```

public void updateUniqueCount(String word)
    throws OperationException {
    // increment the counter for this word
    long newCount = this.wordCountTable.incrementAndGet(Bytes.toBytes(word), 1L);
    if (newCount == 1L) { // first time? Increment unique count
        this.uniqueCountTable.write(new Increment(UNIQUE_COUNT, UNIQUE_COUNT, 1L));
    }
}

```

Note how this method first uses the `incrementAndGet()` method to increase the count for the word, because it needs to know the result to decide whether this is a new unique word. But the second increment is done with the `write()` method of the table, which does not return a result. Unless you really need to get back the result, you should always use that method, because it can be optimized for higher performance by the transaction engine (more details on that in section 0).

Finally, we write a method to retrieve the number of unique words seen. This method is extra cautious as it verifies that the value of the unique count column is actually eight bytes long.

```

public Long readUniqueCount() throws OperationException {
    OperationResult<Map<byte[], byte[]>> result =
        this.uniqueCountTable.read(new Read(UNIQUE_COUNT, UNIQUE_COUNT));
    if (result.isEmpty()) {
        return 0L;
    }
    byte [] countBytes = result.getValue().get(UNIQUE_COUNT);
    if (countBytes == null || countBytes.length != 8) {
        return 0L;
    }
    return Bytes.toLong(countBytes);
}

```

You may have noticed that we only use one single cell of the `uniqueCountTable`, and we could certainly write this code more efficiently (yet the purpose of this example is to illustrate how to implement a custom dataset on top of multiple underlying datasets).

This concludes the implementation of this dataset.

5.5. END-TO-END PROGRAMMING EXAMPLE

To illustrate how all the capabilities of the AppFabric can be put together to build an application, a simple end-to-end example can be implemented using a slightly modified version of the classic Word Count¹.

This application, named **WordCount**, consists of the following:

1. A stream named `wordStream` that receives strings of words to be counted
2. A flow named `WordCounter` that processes the strings from the stream to calculate the word counts and other word statistics using four flowlets:
 1. The `splitter` splits the input string into words and performs any scrubbing.
 2. The `counter` takes words as inputs and performs dataset operations to calculate and persist word count statistics.
 3. The `unique` flowlet is used to calculate the unique number of words seen.
 4. The `associator` stores word associations between all the words in each input string.
3. A procedure named `RetrieveCounts` to serve read requests for the calculated word counts, statistics, and associations. It supports two methods:
 - `getCount` for accessing the word count of a specified word and its word associations.
 - `getStats` for accessing the global word statistics.
4. Four datasets used by the flow and query in order to model, store, and serve the necessary data
 - A core `Table` named `wordStats` to track global word statistics.
 - A system `KeyValueTable` dataset named `wordCounts` to count the occurrences of each word.
 - A custom `UniqueCountTable` dataset named `uniqueCount` to determine and count the number of unique words seen.
 - A custom `AssociationTable` dataset named `wordAssocs` to track associations between words in the input strings.

DEFINING THE APPLICATION

The definition of the application is straightforward and simply wires together all of the different components described:

```
public class WordCount implements Application {
    @Override
    public ApplicationSpecification configure() {
        return ApplicationSpecification.Builder.with()
            .setName("WordCount")
            .setDescription("Example Word Count Application")
            .withStreams()
                .add(new Stream("wordStream"))
            .withDataSets()
                .add(new Table("wordStats"))
                .add(new KeyValueTable("wordCounts"))
                .add(new UniqueCountTable("uniqueCount"))
                .add(new AssociationTable("wordAssocs"))
            .withFlows()
                .add(new WordCounter())
            .withProcedures()
                .add(new RetrieveCounts())
            .build();
    }
}
```

¹ <http://wiki.apache.org/hadoop/WordCount>

```
}  
}
```

DEFINING THE FLOW

The flow must define a `configure` method that wires up the flowlets with their connections. Note that here we need not declare the streams and datasets used:

```
public class WordCounter implements Flow {  
    @Override  
    public FlowSpecification configure() {  
        return FlowSpecification.Builder.with()  
            .setName("WordCounter")  
            .setDescription("Example Word Count Flow")  
            .withFlowlets()  
                .add("splitter", new WordSplitter())  
                .add("counter", new Counter())  
                .add("associator", new WordAssociator())  
                .add("unique", new UniqueCounter())  
            .connect()  
                .fromStream("wordStream").to("splitter")  
                .from("splitter").to("counter")  
                .from("splitter").to("associator")  
                .from("counter").to("unique")  
            .build();  
    }  
}
```

The `splitter` is directly connected to the `wordStream`. It splits each input into words and sends them to the `counter` and the `associator`, the first of which forwards them to the `unique` counter flowlet.

IMPLEMENTING FLOWLETS

With the application and flow defined, it's now time to dig into the actual logic of our application. The processing logic and data write operations occur within flowlets. For each flowlet you extend the `AbstractFlowlet` base class.

The `WordCounter` contains four different flowlets: `splitter`, `counter`, `unique`, and `associator`. The implementation of each of these is shown below with an overview of each.

SPLITTER FLOWLET

`splitter` is the first flowlet in the flow. For each event from the `wordStream`, it interprets the body as a string and splits it into individual words, removing any non-alphabet characters from the words. It then emits each word separately to one of its outputs, for consumption by the `counter`, and the list of all words to its other output, for the `associator`.

```
public class WordSplitter extends AbstractFlowlet {  
  
    @Output("wordOut")  
    private OutputEmitter<String> wordOutput;  
  
    @Output("wordArrayOut")  
    private OutputEmitter<List<String>> wordListOutput;  
  
    public void process(StreamEvent event) {  
        // input is a string, need to split it by whitespace  
    }  
}
```

```

byte [] rawInput = Bytes.toBytes(event.getBody());
String inputString = new String(rawInput);

String [] words = inputString.split("\\s+");
List<String> wordList = new ArrayList<String>(words.length);

// We have an array of words, now remove all non-alpha characters
for (String word : words) {
    word = word.replaceAll("[^A-Za-z]", "");
    if (!word.isEmpty()) {
        // emit every word that remains to the counter
        wordOutput.emit(word);
        wordList.add(word);
    }
}
// Send the list of words to the associater
wordListOutput.emit(wordList);
}
}

```

COUNTER FLOWLET

The `counter` flowlet receives single words as inputs. It counts the number of occurrences for each word in the key/value table `wordCounts`, and keeps track of the total word and character count in the core table `wordStats`:

```

public class Counter extends AbstractFlowlet {

    @UseDataSet("wordStats")
    private Table wordStatsTable;

    @UseDataSet("wordCounts")
    private KeyValueTable wordCountsTable;

    byte[] TOTALS_ROW = Bytes.toBytes("totals");
    byte[] TOTAL_LENGTH = Bytes.toBytes("total_length");
    byte[] TOTAL_WORDS = Bytes.toBytes("total_words");

    private OutputEmitter<String> wordOutput;

    @ProcessInput("wordOut")
    public void process(String word) throws OperationException {
        // Count number of times we have seen this word
        this.wordCountsTable.increment(Bytes.toBytes(word), 1L);

        // Count other word statistics (word length, total words seen)
        this.wordStatsTable.write(
            new Increment(TOTALS_ROW,
                new byte[][] { TOTAL_LENGTH, TOTAL_WORDS },
                new long[] { word.length(), 1L}));

        // Forward the word to the unique counter flowlet
        wordOutput.emit(word);
    }
}

```

UNIQUE COUNTER FLOWLET

The `unique` flowlet receives each word from the `counter` flowlet. All its data logic is coded into the `UniqueCountTable` dataset – we already know it from Section 0. This dataset uses tables to determine the number of unique words seen.

```
public class UniqueCounter extends AbstractFlowlet {

    @UseDataSet("uniqueCount")
    private UniqueCountTable uniqueCountTable;

    public void process(String word) throws OperationException {
        this.uniqueCountTable.updateUniqueCount(word);
    }
}
```

ASSOCIATOR FLOWLET

The `associator` flowlet receives arrays of words and stores associations between them using a custom dataset, the `AssociationTable`, described below.

```
public class WordAssociator extends AbstractFlowlet {

    @UseDataSet("wordAssocs")
    private AssociationTable associationTable;

    public void process(String [] words) throws OperationException {
        // Store word associations
        Set<String> wordSet = new TreeSet<String>(Arrays.asList(words));
        this.associationTable.writeWordAssocs(wordSet);
    }
}
```

Note that even though `process` expects a set of strings, it can consume lists of strings from the `counter` flowlet – that is the magic of type projection!

IMPLEMENTING CUSTOM DATASETS

This application uses four datasets: A core table, a system key/value table, and two custom datasets built using core tables. The first custom dataset is the `UniqueCountTable` that we already discussed in section 0.

The other custom dataset is the `AssociationTable` and it tracks associations between words by counting the number of times they occur together. Rather than requiring that this pattern be implemented within our flowlets and queries, it is implemented as a custom dataset, exposing a simple and specific API rather than a complex and generic one. Its interface has two public methods, `writeWordAssocs()` and `readWordAssocs()` that both use a core table. First of all, as all datasets, it must define two constructors and a `configure()` method (see section 0):

```
public class AssociationTable extends DataSet {

    private Table table;

    public AssociationTable(String name) {
        super(name);
        this.table = new Table("word_assoc_" + name);
    }
}
```

```

public AssociationTable(DataSetSpecification spec) {
    super(spec);
    this.table = new Table(
        spec.getSpecificationFor("word_assoc_" + this.getName()));
}

@Override
public DataSetSpecification configure() {
    return new DataSetSpecification.Builder(this)
        .dataset(this.table.configure())
        .create();
}

```

The dataset operates on bags of words to compute for each word the set of other words that most frequently occur in the same bag. It uses a columnar table to count the number of times that two words occur together. That counter is in a cell of the table with the first word as the row key and the second word as the column key.

```

public void writeWordAssocs(Set<String> words) throws OperationException {
    for (String rootWord : words) {
        for (String assocWord : words) {
            if (!rootWord.equals(assocWord)) {
                this.table.write(new Increment(Bytes.toBytes(rootWord),
                    Bytes.toBytes(assocWord), 1L));
            }
        }
    }
}

```

Note that this table is very sparse – most words never occur together and will never be counted. In a table with a fixed schema, such as a relational table, this would be a very space-consuming representation. In a columnar table, however, non-existent cells occupy (almost) no space. Furthermore, we can use the second word, which is only known at runtime, as the column key. That would also be impossible in a traditional relational table.

To read the most frequent associations for a word, the dataset method simply reads the row for the word, iterates over all columns and passes them to a top-K collector (let's assume that is already implemented).

```

public Map<String,Long> readWordAssocs(String word, int limit)
    throws OperationException {

    // Retrieve all columns of the word's row
    OperationResult<Map<byte[], byte[]>> result =
        this.table.read(new Read(Bytes.toBytes(word), null, null));
    TopKCollector collector = new TopKCollector(limit);
    if (!result.isEmpty()) {
        // iterate over all columns
        for (Map.Entry<byte[],byte[]> entry : result.getValue().entrySet()) {
            collector.add(Bytes.toLong(entry.getValue()),
                Bytes.toString(entry.getKey()));
        }
    }
    return collector.getTopK();
}

```

IMPLEMENTING A PROCEDURE

To read the data written by the flow, we implement a procedure, which will bind a handler to a REST endpoint to get external access. The procedure has two methods, one to get the overall statistics, and one to get the statistics for a single word. It begins with the `configure` method that all procedures must have. It also declares the datasets that it uses, namely all four datasets of the application.

```
public class RetrieveCounts extends AbstractProcedure {

    @Override
    public ProcedureSpecification configure() {
        return ProcedureSpecification.Builder.with()
            .setName("RetrieveCount")
            .setDescription("Example Word Count Procedure")
            .build();
    }

    @UseDataSet("wordStats")
    private Table wordStatsTable;

    @UseDataSet("wordCounts")
    private KeyValueTable wordCountsTable;

    @UseDataSet("uniqueCount")
    private UniqueCountTable uniqueCountTable;

    @UseDataSet("wordAssocs")
    private AssociationTable associationTable;
```

Now we can implement a handler for the first method, `getStats`. It return general statistics across all words seen:

```
@Handle("getStats")
public void getStats(ProcedureRequest request,
                    ProcedureResponder responder) throws Exception {

    long totalWords = 0L, uniqueWords = 0L;
    double averageLength = 0.0;

    // Read the total length and total count to calculate average length
    OperationResult<Map<byte[],byte[]>> result =
        this.wordStatsTable.read(
            new Read(TOTALS_ROW, new byte[][] { TOTAL_LENGTH, TOTAL_WORDS }));
    if (!result.isEmpty()) {
        // extract the total sum of lengths
        byte[] lengthBytes = result.getValue().get(TOTAL_LENGTH);
        Long totalLength = lengthBytes == null ? 0L : Bytes.toLong(lengthBytes);
        // extract the total count of words
        byte[] wordsBytes = result.getValue().get(TOTAL_WORDS);
        totalWords = wordsBytes == null ? 0L : Bytes.toLong(wordsBytes);
        // compute the average length
        if (totalLength != 0 && totalWords != 0) {
            averageLength = (double)totalLength/(double)totalWords;
            // Read the unique word count
            uniqueWords = this.uniqueCountTable.readUniqueCount();
        }
    }
}
```



```

// return a map as JSON
Map<String, Object> results = new TreeMap<String, Object>();
results.put("totalWords", totalWords);
results.put("uniqueWords", uniqueWords);
results.put("averageLength", averageLength);
responder.sendJson(new ProcedureResponse(Code.SUCCESS), results);
}

```

The second handler is for the method `getCount`. Given a word, it returns the count of the word together with the top words associated with that word, up to a specified limit, or up to 10 if no limit is given.

```

@Handle("getCount")
public void getCount(ProcedureRequest request,
                    ProcedureResponder responder) throws Exception {
    String word = request.getArgument("word");
    if (word == null) {
        responder.error(Code.CLIENT_ERROR,
            "Method 'getCount' requires argument 'word'");
        return;
    }

    String limitArg = request.getArgument("limit");
    int limit = limitArg == null ? 10 : Integer.valueOf(limitArg);

    // Read the word count
    byte[] countBytes = this.wordCountsTable.read(Bytes.toBytes(word));
    Long wordCount = countBytes == null ? 0L : Bytes.toLong(countBytes);

    // Read the top associated words
    Map<String, Long> wordsAssocs =
        this.associationTable.readWordAssocs(word, limit);

    // return a map as JSON
    Map<String, Object> results = new TreeMap<String, Object>();
    results.put("word", word);
    results.put("count", wordCount);
    results.put("assocs", wordsAssocs);
    responder.sendJson(new ProcedureResponse(Code.SUCCESS), results);
}

```

This concludes the `WordCount` example. You can find it in the examples directory of the Developer Suite.

5.6. TESTING YOUR APPLICATION

The AppFabric comes with a convenient way to unit test your application. The base for such a test is the `AppFabricTestBase`, which is packaged separately from the API in its own artifact, because it depends on all the runtime classes of the AppFabric. You can include it into your test dependencies in two ways:

- Include all jar files in the `lib` directory of the Developer Suite installation
- Include the `continuity-test` artifact in your Maven test dependencies (see the `pom.xml` of `WordCount` for an example).

Note that for building an application, you only need to include the AppFabric API in your dependencies; for testing, however, you need the AppFabric run-time. To build your test case, you extend the `AppFabricTestBase` class. Let us write a test case for the `WordCount` example:

```
public class WordCountTest extends AppFabricTestBase {  
  
    @Test  
    public void testWordCount() throws Exception {
```

The first thing we do in this test is to deploy the application. Then we can start the flow and the procedure:

```
// deploy the application  
ApplicationManager appManager = deployApplication(WordCount.class);  
  
// start the flow and the procedure  
FlowManager flowManager = appManager.startFlow("WordCounter");  
ProcedureManager procManager = appManager.startProcedure("RetrieveCount");
```

Now that the flow is running, we can send some events to the stream:

```
// send a few events to the stream  
StreamWriter writer = appManager.getStreamWriter("wordStream");  
writer.send("hello world");  
writer.send("a wonderful world");  
writer.send("the world says hello");
```

To wait for all events to be processed, we can get a metrics observer for the last flowlet in the pipeline, the word associator, and wait for its processed count to reach 3, or time out after 5 seconds:

```
// wait for the events to be processed, or at most 5 seconds  
RuntimeMetrics metrics = RuntimeStats.  
    getFlowletMetrics("WordCount", "WordCounter", "associator");  
metrics.waitForProcessed(3, 5, TimeUnit.SECONDS);
```

Now we can start verifying that the processing was correct. Start by obtaining a client for the procedure, and submitting a query for the global statistics:

```
// now call the procedure  
ProcedureClient client = procManager.getClient();  
// first verify global statistics  
String response = client.query("getStats", Collections.EMPTY_MAP);
```

If the query fails for any reason, then this method would throw an exception. In case of success, the response is a JSON string. We must deserialize the JSON in order to verify the results:

```

Map<String, String> map = new Gson().fromJson(response, stringMapType);
Assert.assertEquals("9", map.get("totalWords"));
Assert.assertEquals("6", map.get("uniqueWords"));
Assert.assertEquals(((double)42)/9,
    (double)Double.valueOf(map.get("averageLength")), 0.001);

```

Then we ask for the statistics of one of the words in the test events. The verification is a little more complex, because we now have a nested map as a response, and the value types in the top-level map are not uniform.

```

// now verify some statistics for one of the words
response = client.query("getCount", ImmutableMap.of("word","world"));
Map<String, Object> omap = new Gson().fromJson(response, objectMapType);
Assert.assertEquals("world", omap.get("word"));
Assert.assertEquals(3.0, omap.get("count"));
// the associations are a map within the map
Map<String, Double> assocs = (Map<String, Double>) omap.get("assocs");
Assert.assertEquals(2.0, (double)assocs.get("hello"), 0.000001);
Assert.assertTrue(assocs.containsKey("hello"));
}
}

```

6. API AND TOOL REFERENCE

6.1. JAVA APIS

The Javadoc for all Core AppFabric Java APIs is included in the Developer Suite:

[./continuity-developer-edition-1.4.0/docs/javadoc/index.html](http://continuity-developer-edition-1.4.0/docs/javadoc/index.html)

Note that some APIs are annotated as **@Beta**. They represent experimental features that have not been fully tested nor documented yet – they may or may not be functional. Also, these APIs may be removed in future versions of the AppFabric SDK. Use them at your own discretion.

6.2. REST APIS

The Continuity AppFabric Platform exposes these REST interfaces:

1. **Stream:** To send data events to a stream or to inspect the contents of a stream.
2. **Data:** To clean up the Data Fabric
3. **Procedure:** To send queries to a procedure.
4. **Monitor:** To monitor the status of running applications and to retrieve metrics.

Common return codes for all REST calls:

- *200 OK:* The request returned successfully
- *400 Bad Request:* The request had a combination of parameters that is not recognized/allowed
- *401 Unauthorized:* The request did not contain an authentication token
- *403 Not Allowed:* The request was authenticated but the client does not have permission
- *404 Not Found:* The request did not address any of the known URIs
- *405 Method Not Allowed:* A request with an unsupported method was received
- *500 Internal Server Error:* An internal error occurred while processing the request
- *501 Not Implemented:* A request contained a query, which is not supported by this API

Note: These may be omitted from the descriptions below but any request may return them.

REST ENDPOINT PORT CONFIGURATION

By default, each of the three REST interfaces binds to a set of default ports on localhost. You can modify these ports by modifying your *continuity-site.xml* within your Developer Suite [conf/](#) directory. The properties you will modify are:

Description	Property	Default Value
Stream REST Port	stream.rest.port	10000
Data REST Port	data.rest.port	10002
Procedure REST Port	procedure.rest.port	10010
Monitor REST Port	monitor.rest.port	10005

The descriptions below will assume the default ports.

When you interact with a developer sandbox, all REST APIs require that you to use SSL for the connection, and that you authenticate your request by sending your API key in an HTTP header:

X-Continuity-APIKey : <API key>

STREAM REST API

This interface allows creating streams, sending events to a stream and reading single events from a stream.

CREATING A STREAM

A stream can be created with an HTTP put request:

```
PUT http://<hostname>:10000/stream/<new-stream-id>
```

The request returns *200 OK* in case of success.

SENDING EVENTS

A request to send an event to a stream is an HTTP post. The URI is formed as

```
POST http://<hostname>:10000/stream/<stream-id>
```

where the *<stream-id>* identifies an existing stream. The body of the request must contain the event in binary form. You can pass headers for the event as HTTP headers, by prefixing them with the stream id, that is:

```
<stream-id>.<property> : <string value>
```

After receiving the request, the REST connector will transform it into a stream event as follows:

- The body of the event is an identical copy of the bytes in the body of the HTTP post request
- If the request contains any headers prefixed with the stream id, then stream id prefix is stripped from the header name and the header is added to the event.

Return codes for the request are:

- *200 OK*: Everything went well.
- *404 Not found*: The stream does not exist.

The response will always have an empty body.

READING EVENTS

Streams may have multiple consumers (e.g. multiple flows), each of which may be a group of different agents (e.g. multiple instances of a flowlet). In order to read, a client must first obtain a consumer (group) id, which needs to be passed to subsequent read requests.

Getting a consumer id is performed as an HTTP GET to the URL

```
GET http://<hostname>:10000/stream/<stream-id>?q=newConsumer
```

The new consumer id is returned in a response header and, for convenience, also in the body of the response.

```
X-Continuity-ConsumerId: <consumer-id>
```

Return codes:

- *201 Created*: Everything went well, and the new consumer is returned.
- *404 Not found*: The stream does not exist.

Once this is completed single events can be read from the stream, in the same way that a flow reads events. That is, the read will always return the event from the queue that was inserted first and has not been read yet (FIFO semantics). In order to read the third event that was sent to a stream, two previous reads have to be performed.

Note that you can always start reading from the first event by getting a new consumer id. A read is performed as an HTTP get to the URI:

GET http://<hostname>:10000/stream/<stream-id>?q=dequeue

and the request must pass the consumer id in a header of the form

X-Continuity-ConsumerId: <consumer-id>

The response will contain the binary body of the event in its body and a header

<stream-id>.<property> : <value>

for each header of the stream event, analogous to how you send headers in the post request.

Return codes:

- *200 OK*: Everything went well.
- *204 No Content*: The stream exists but it is empty or the given consumer id has read all events in the stream.
- *404 Not found*: The stream does not exist.

READING MULTIPLE EVENTS

Reading multiple events is not supported directly by the stream API, but the stream-client tool has a way to view all, the first N, or the last N events in the stream. See the command-line guide in Section 6.3.

DATA REST API

The data API allows you to interact with tables (the core datasets) through REST: You can create tables and read, write, modify, or delete data.

CREATING A TABLE

To create a new table, you issue an HTTP put request:

```
PUT http://<hostname>:10002/data/Table/<table-name>
```

This will create a table with the given name. The table name should only contain ASCII letters, digits and hyphens. If a table with the same name already exists, no error is returned, and the existing table remains in place. However, if a dataset of a different type exists with the same name, for example a key/value table, this call will return an error. The expected return codes are:

- *200 OK*: Everything went well.
- *409 Conflict*: A dataset of a different type already exists with the given name.

WRITING DATA TO A TABLE

To write to a table, you send an HTTP put request to the table's URL:

```
PUT http://<hostname>:10002/data/Table/<table-name>/<row-key>
```

In the body, you must specify the columns and values that you want to write as a JSON string map, for example:

```
{"x":"y","y":"a","z":"1"}
```

This writes three columns named *x*, *y*, and *z* with values *y*, *a*, and *1*, respectively. Note that we specify all column keys and values are strings, whereas in the Data Fabric, all keys and values are byte arrays. For now, let us just assume that these byte values can simply be converted to Strings, and we will soon learn how to use keys and values that are not ASCII strings.

The request will return:

- *200 OK*: Everything went well.
- *400 Bad Request*: The JSON string is not well-formed or cannot be parsed as a map from string to string.
- *404 Not found*: A table with the given name does not exist.

READING DATA FROM A TABLE

To read from a table, you address the row that you want to read directly in an HTTP get request:

```
GET http://<hostname>:10002/data/Table/<table-name>/<row-key>
```

The response will be a JSON string representing a map from column name to value. For example, reading back the row that we wrote in the previous section, the response is:

```
{"x":"y","y":"a","z":"1"}
```

If you are only interested in some of the columns, you can specify a list of columns explicitly or give a range of columns, in all the same ways that you specify the columns for a [Read](#) operation (see Section 0, Core Datasets - Tables). For example:

<code>GET ...<table-name>/<row-key>?columns=x,y</code>	returns only columns <code>x</code> and <code>y</code> .
<code>GET ...<table-name>/<row-key>?start=c5</code>	returns all columns greater or equal to <code>c5</code> .
<code>GET ...<table-name>/<row-key>?stop=c5</code>	returns all columns less than (exclusive) <code>c5</code> .
<code>GET ...<table-name>/<row-key>?start=c2&stop=c5</code>	returns all columns between <code>c2</code> and (exclusive) <code>c5</code> .

The request will return:

- `200 OK`: Everything went well.
- `404 Not found`: A table with the given name does not exist.

INCREMENTING DATA IN A TABLE

You can also perform an atomic increment of cells of a table, and receive back the incremented values, by posting to the following URL (note that incrementing is not idempotent and hence cannot be an HTTP put).

`POST http://<hostname>:10002/data/Table/<table-name>/<row-key>?op=increment`

In the body, you must specify the columns and values that you want to write as a JSON map from strings to long numbers, for example:

```
{"x":1,"y":7}
```

This REST call has the same effect as the corresponding table `Increment` operation (see Section 0, Core Datasets - Tables). If successful, the response contains a JSON map from column key to the incremented values. For example, the existing value of column `x` was 4, and column `y` did not exist, then the response is (column `y` is newly created):

```
{"x":5,"y":7}
```

The expected HTTP return codes are:

- `200 OK`: Everything went well.
- `400 Bad Request`: The JSON string is not well-formed or cannot be parsed as a map from string to long, or one of the existing column values is not an 8-byte long value.
- `404 Not found`: A table with the given name does not exist.

DELETING DATA FROM A TABLE

To delete from a table, you submit an HTTP delete request:

`DELETE http://<hostname>:10002/data/Table/<table-name>/<row-key>?columns=<column-key,...>`

You must explicitly list the columns that you want to delete. The expected return codes are:

- `200 OK`: Everything went well.
- `400`: No columns were specified.
- `404 Not found`: A table with the given name does not exist.

ENCODING OF KEYS OF VALUES

The URLs and JSON bodies of your REST requests contain row keys, column keys and values, all of which are binary byte arrays in the Java API (see Section 0, Core Datasets - Tables). Therefore you need a way to encode binary keys and values as strings in the URL and the JSON body. The `encoding` parameter of the URL specifies this encoding. For example, if you append a parameter `encoding=hex` to the request URL, then all keys and values are interpreted

as hexadecimal strings, and returned JSON from read and increment requests also has the keys and values encoded that way. But be aware that this applies to all keys and values involved in the request: Suppose you incremented a column in a new table by 42:

```
POST http://<hostname>:10002/data/Table/counters/a?op=increment
```

```
{ "x" : 42 }
```

Now the value of column `x` is the 8-byte number 42. If you query for the value of this with

```
GET http://<hostname>:10002/data/Table/counters/a?columns=x
```

Then the returned JSON will contain a non-printable string for the value of column `x`:

```
{ "x": "\u0000\u0000\u0000\u0000\u0000\u0000\u0000\u0000*" }
```

Note the Unicode escapes in the string, and the asterisk at the end (which happens to be the character at code point 42). To make this more legible, you can request for hexadecimal notation, and that will require that you also encode the row key and the column key in your request as hexadecimal:

```
GET http://<hostname>:10002/data/Table/counters/61?columns=78
```

The response now contains both the column key and the value as hexadecimal strings.

```
{ "78": "000000000000002a" }
```

The supported encodings are:

- Default. Only ASCII characters are supported and mapped to bytes one-to-one.
- `encoding=hex` Hexadecimal strings. For example, the ASCII string `a:b` is represented as `613A62`.
- `encoding=url` URL encoding (also known as %-encoding). URL-safe characters use ASCII-encoding, other bytes values are escaped using a % sign. For example, the hexadecimal value `613A62` is represented as the string `a%3Ab`.
- `encoding=base64` URL-safe Base-64 encoding without padding. For more information, see [Internet RFC 2045](#). For example, the hexadecimal value `613A62` is represented as the string `YTpi`.

If you specify an encoding that is not supported, or you specify keys or values that cannot be decoded using the that encoding, the request will return HTTP code `400 Bad Request`.

PROCEDURE REST API

This interface allows sending queries to the procedures of your application.

EXECUTING QUERY

Remember that a procedure accepts a method name and a map of string arguments as parameters. To send a query to a procedure, you send the method name as part of the request URI and the arguments as a JSON string in the body of the request. The request is an HTTP post:

```
POST http://<hostname>:10010/query/<procedure>/<method>
```

For example, to invoke the `getCount` method of the `RetrieveCounts` procedure from Section 5.5, you post:

```
POST http://<hostname>:10010/query/RetrieveCounts/getCount
```

```
{ "word" : "a" }
```

Return codes:

- *200 OK*: Everything went well, and the body contains the query result.
- *400 Bad Request*: The procedure and method exist, but the arguments are not as expected
- *404 Not found*: The procedure or the method does not exist.

MONITOR REST API

The monitor API lets you get the status and metrics about running flows and procedures. To get the status of a running flow, you submit an HTTP get request:

```
GET http://<hostname>:10005/monitor/<application>/<flow>/status
```

The same works for procedures, just use the procedure name instead of the flow name in the URI. To get metrics about the flow, you change the request slightly:

```
GET http://<hostname>:10005/monitor/<application>/<flow>/metrics
```

This will return a comma-separated list of metric names with their values. Note that metrics are only generated by a running flow or procedure. If a flow has never run, or a procedure has never received any requests, the list of metrics is empty.

Return codes:

- *200 OK*: Everything went well, and the body contains the requested status or metrics.
- *404 Not found*: The procedure or the flow does not exist.

6.3. COMMAND LINE TOOLS

The Developer Suite includes a suite of tools that allow you to access and manage local and remote AppFabric instances from the command line. The list of tools is outlined below. They all support the `--help` option to get brief usage information.

The examples in the rest of this section assume you are in the `bin` directory of the Developer Suite (e.g. `~/continuity-developer-suite-1.3.0/bin/`)

APPFABRIC

The AppFabric shell script can be used to start and stop the AppFabric server:

```
> continuity-app-fabric start
> continuity-app-fabric stop
```

To get usage information for the tool invoke it with the `--help` parameter:

```
> continuity-app-fabric --help
```

DATA CLIENT

The data client command line tool can be used to create tables and to read, write, modify, or delete data in tables.

To create a table, you invoke:

```
> data-client create --table myTable
```

To write data to the table, you specify the row key, column keys and values on the command line:

```
> data-client write --table myTable --row a --column x --value z
> data-client write --table myTable --row a --column x --value z --column y --value q
> data-client write --table myTable --row a --columns x,y --values z,q
```

To read from a table, you can read the entire row, specific columns, or a column range (compare Section 0, Core Datasets - Tables):

```
> data-client read --table myTable --row a
> data-client read --table myTable --row a --column x
> data-client read --table myTable --row a --columns x,y
> data-client read --table myTable --row a --start x
> data-client read --table myTable --row a --stop z
> data-client read --table myTable --row a --start y --stop z
```

The `read` command prints the columns that it retrieved to the console:

```
> data-client read --table myTable --row a
x:z
y:q
```

If you prefer JSON output, you can use the `--json` argument:

```
> data-client read --table myTable --row a --json
{"x":"z","y":"q"}
```

To delete from a table, you specify the row and the columns to delete on the command line:

```
> data-client delete --table myTable --row a --columns=x,y
```

You can also perform atomic increment on cells of a table. The command prints the incremented values on the console:

```
> data-client increment --table myTable --row counts --columns x,y --values 1,4
x:1
y:4
> data-client increment --table myTable --row counts --columns x,y --values 2,-6
x:3
y:-2
```

Similarly to the REST interface, the command line allows to use an encoding for binary values or keys. If you specify an encoding, then it applies to all keys and values involved in the command. For example, the following command has the same effect as the REST call on in Section 6.2 (Encoding of Keys of Values):

```
> data-client read --table counters --row 61 --hex
78:000000000000002a
```

Other supported encodings are URL-safe Base-64 with `--base64` and URL-encoding ("`%-escaping`") with `-url`. See Section 6.2 (Encoding of Keys of Values) for more details.

To use the data-client with your developer sandbox, you need to provide the host name of your sandbox and the API key that authenticates you with the sandbox:

```
> data-client create --table myTable --host <hostname> --apikey <apikey>
```

If you configured your local AppFabric to use different REST ports (see Section 6.2), then you also need to specify the default data REST port on the command-line:

```
> data-client create --table myTable --host <hostname> --apikey <apikey> --port 10002
```

STREAM CLIENT

The stream client is a utility to send events to a stream or to view the current content of a stream. To send a single event to a stream, you invoke:

```
> stream-client send --stream text --header number "101" --body "message 101"
```

The stream must already exist when you submit this. The send command supports adding multiple headers:

```
> stream-client send --stream text --header number "102" --header category "info"
>> --body "message 102"
```

Since the body of an event is binary, it is not always printable text. You can use the `--hex` option to specify body in hexadecimal (the default is URL-encoding). If the body is too long or too inconvenient to specify it on the command line, you can use `--body-file <filename>` as an alternative to `--body` to read it from a binary file.

To inspect the contents of a stream, you can use the `view` command:

```
> stream-client view --stream msgStream --last 50
```

This retrieves and prints the last (that is, the latest) 50 events from the stream. Alternatively, you can use `-first` to see the first (oldest) events, or `--all` to see all events in the stream.

As with the send command, you can use `--hex` to print the body of each event in hexadecimal form. Also, similar to the data client (see the previous section), you can use the `--host`, `--port`, and `--apikey` options to use the stream client with your developer sandbox (the default stream REST port is 10000):

```
> stream-client view --stream text --host <hostname> --apikey <apikey> --port 10000
```

In order to create a stream that does not exist yet, invoke:

```
> stream-client create --stream newStream
```

APPFABRIC CLIENT

The AppFabric command line client lets you deploy applications, start and stop flows and procedures, and promote applications to a remote AppFabric instance, that is, your developer sandbox.

To deploy an application that is already packaged into an archive file in your local file system, you invoke:

```
> app-fabric-client deploy --archive myApp.jar
```

You also invoke the same command to update an application to a newer version. Be aware that all flows and procedures of the application must be stopped in order to perform the update.

After the application is deployed, you can start and stop a flow or a procedure, or get their status:

```
> app-fabric-client start --application myapp --flow myflow  
> app-fabric-client stop --application myapp --flow myproc  
> app-fabric-client status --application myapp --flow myflow
```

If you are ready to promote your application to the developer sandbox (Push-to-Cloud), you need the hostname of your sandbox and the API key that authenticates you with the sandbox:

```
> app-fabric-client promote --application myapp --host <hostname> --apikey <apikey>
```

6.4. ECLIPSE IDE PLUGIN

This section walks through creating, running, and debugging a simple application using the Continuity Eclipse IDE plugin.

PREREQUISITES

1. Eclipse Juno (available at <http://eclipse.org/downloads/>)
2. The Continuity Developer Suite (available from <https://accounts.continuity.com/>)
3. The Continuity Eclipse Plugin, packaged with the Developer Suite as a jar file named *continuity-eclipse-plugin-1.4.0.jar*

GETTING ECLIPSE AND SETTING UP PLUGIN

1. Download and unpack the Eclipse IDE.
2. Install the Continuity Eclipse plugin by copying the plugin jar into the “plugins” subdirectory of the unpacked Eclipse directory. For example:

```
> cp ~/continuity-developer-edition-1.4.0/continuity-eclipse-plugin-1.4.0.jar  
$ECLIPSE_HOME/eclipse/plugins/
```

3. Start Eclipse.

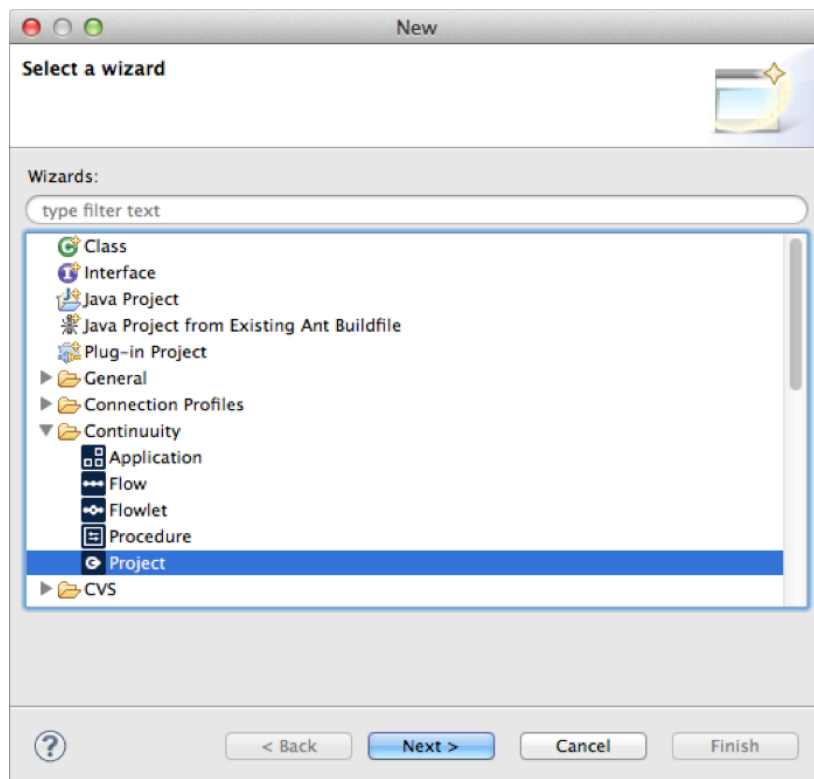
CREATING A SIMPLE APPLICATION

This section guides you through creating a simple application from scratch in Eclipse IDE.

CREATE A NEW CONTINUITY PROJECT

1. Create a new project by following the path

File -> New -> Other... -> Continuity -> Project



NOTE: if you don't see the "Continuity" menu item in this dialog, the plugin was not installed correctly. Please refer to the first section of the document for the installation steps.

2. Enter the name of your application, e.g. "FooApplication" and click Next.

Create a Continuity Project
Enter Continuity Project name

Project name:

☒ Use default location

Location: [Browse...](#)

JRE

☒ Use an execution environment JRE:

☐ Use a project specific JRE:

☐ Use default JRE (currently 'Java SE 6 (MacOS X Default)') [Configure JREs...](#)

Project layout

☐ Use project folder as root for sources and class files

☒ Create separate folders for sources and class files [Configure default...](#)

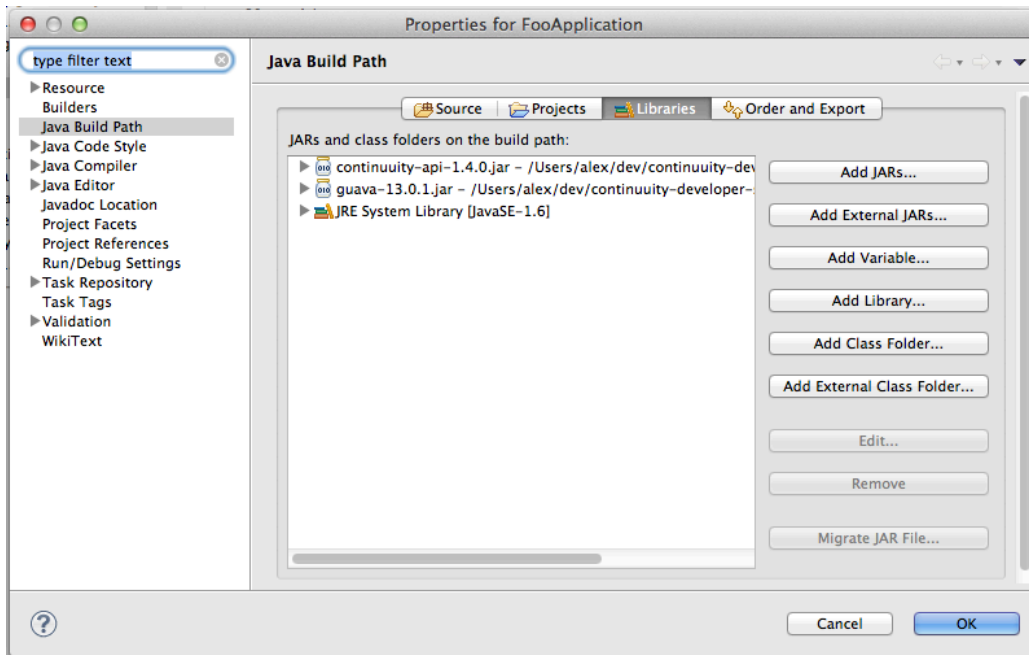
Working sets

☐ Add project to working sets

Working sets: [Select...](#)

[?](#) [< Back](#) [Next >](#) [Cancel](#) [Finish](#)

3. Add "continuity-api-1.4.0.jar" into the class path by using the standard "Add External JARs..." dialog on the "Libraries" tab. The file is located under <home>/continuity-developer-edition-1.4.0.
4. Add "guava-13.0.1.jar" located under <home>/continuity-developer-edition-1.4.0/lib using the same mechanism.



CREATING A FLOW

"FooApplication" will contain a simple Flow called "MyFlow". This Flow will contain a single Flowlet called "MyFlowlet" that will consume data from a Stream called "inStr".

ADD A FLOW AND FLOWLET CLASSES

Create a Flow class stub:

1. Use the Eclipse File menu. Go to:
File -> New -> Other... -> Continuity -> Flow
2. Enter package name "my" and class name "MyFlow".

Create a Flowlet class stub:

1. Use the Eclipse File menu. Go to:
New -> Other... -> Continuity -> Flowlet
2. Enter package name "my" and class name "MyFlowlet"

CONFIGURE FLOW

The flow will consist of one Flowlet that consumes data from a Stream and outputs consumed data to the output console. First, we need to configure the Flow.

Go to my/MyFlow.java add the following to *configure()* :

```
return FlowSpecification.Builder.with()
    .setName("MyFlow")
    .setDescription("MyFlow Description")
    .withFlowlets().add(new MyFlowlet())
    .connect().fromStream("inStr").to("MyFlowlet")
    .build();
```


- “inStr” is the name of the stream that is consumed by the Flowlet.
- “MyFlowlet” is the name of the Flowlet
- The stream is connected to the Flowlet using connect() followed by the *.fromStream().to()* syntax

NOTE: you can use any names instead of the ones in the example. This defines the identity of the flow and how streams and flowlet are interconnected. The resulting flow code should contain the following:

```
package my;
import com.continuity.api.flow.Flow;
import com.continuity.api.flow.FlowSpecification;

public class MyFlow implements Flow {
    @Override
    public FlowSpecification configure() {
        return FlowSpecification.Builder.with()
            .setName("MyFlow")
            .setDescription("Flow description")
            .withFlowlets().add(new MyFlowlet())
            .connect().fromStream("inStr").to("MyFlowlet")
            .build();
    }
}
```

CONFIGURING FLOWLET AND ADDING PROCESSING LOGIC

The Flowlet will consume incoming data from a stream and output incoming data to the output console. To configure the flowlet:

```
@Override
public FlowletSpecification configure() {
    return FlowletSpecification.Builder.with()
        .setName("MyFlowlet")
        .setDescription("flowlet description")
        .build();
}
```

To do this, go to my/MyFlowlet.java in the editor and add to process method:

```
@ProcessInput("inStr")
public void process(StreamEvent event) {
    System.out.println("Input value: " +
        new String(Bytes.toBytes(event.getBody())));
}
```

The final flowlet code should contain the following:

```
package my;
import com.continuity.api.annotation.*;
import com.continuity.api.common.Bytes;
import com.continuity.api.flow.flowlet.*;

public class MyFlowlet extends AbstractFlowlet {

    @Override
    public void initialize(FlowletContext context) throws FlowletException {
        // TODO Auto-generated method stub
    }

    @Override
    public FlowletSpecification configure() {
        return FlowletSpecification.Builder.with()
            .setName("MyFlowlet")
            .setDescription("flowlet description")
            .build();
    }

    @Override
    public void destroy() {
        // TODO Auto-generated method stub
    }

    @ProcessInput("inStr")
    public void process(StreamEvent event) {
        System.out.println("Input value: " +
            new String(Bytes.toBytes(event.getBody())));
    }
}
```

CONFIGURE THE APPLICATION

Finally, create an application that ties all of the flowlets together. To create an Application, use the Eclipse File menu. Go to:

File -> New -> Other... -> Continuity -> Application

In the opened dialog enter the package and class names, e.g. package "my" and class name "FooApplication".

CONFIGURE APPLICATION

Add the following line to the configure() method in the just-created Application class –

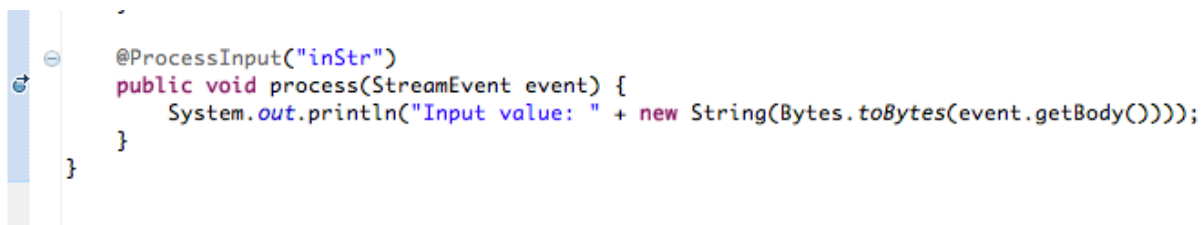
```
@Override
public ApplicationSpecification configure() {
    return ApplicationSpecification.Builder.with()
        .setName("FooApplication")
        .setDescription("FooApplication Description")
        .withStreams().add(new Stream("inStr"))
        .noDataSet()
        .withFlows().add(new MyFlow())
        .noProcedure().build();
}
```

RUNNING AND DEBUGGING APPLICATION

This section guides through running and debugging steps for the application created above.

SET BREAKPOINT

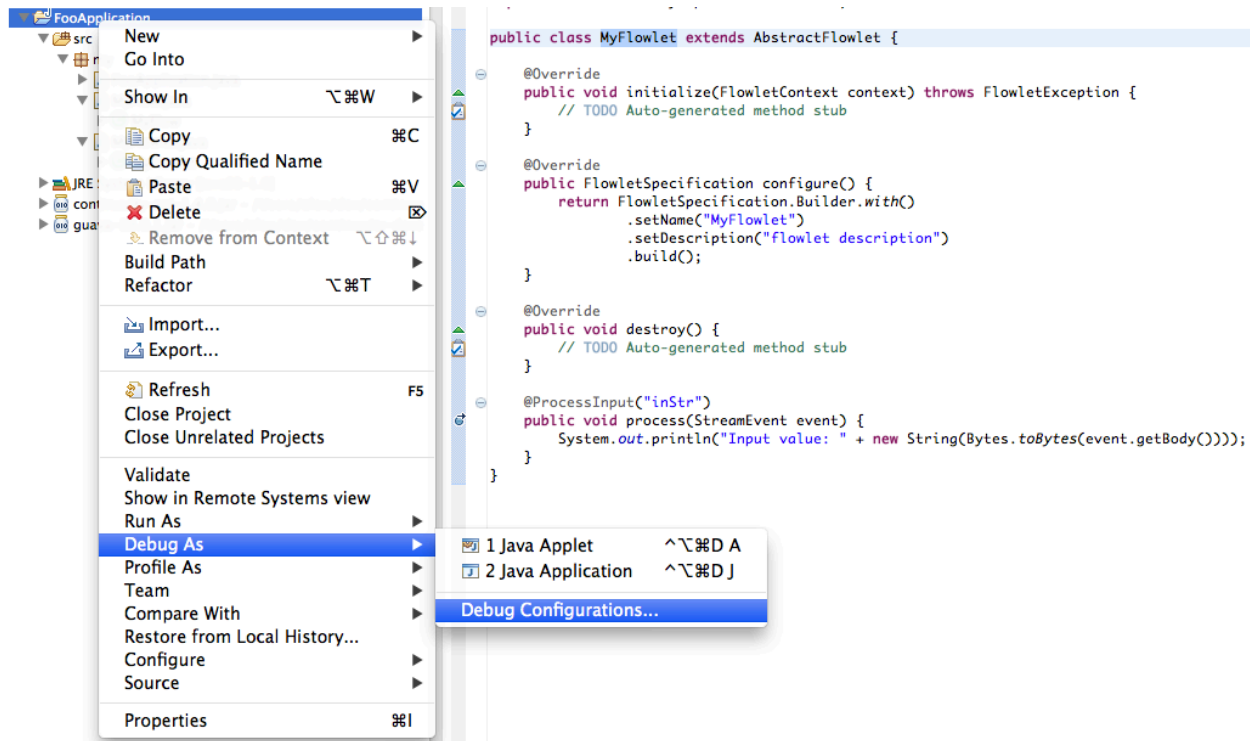
We can set breakpoints anywhere in application code before debugging. E.g. it may be useful to inspect execution of the MyFowlet.proces() method:

A screenshot of the Eclipse IDE showing a code editor with a Java class. A breakpoint is set on the line `public void process(StreamEvent event) {`. The code includes an annotation `@ProcessInput("inStr")` and a call to `System.out.println` that prints the input value. The IDE interface shows a sidebar on the left with a breakpoint icon and a line number column.

```
@ProcessInput("inStr")
public void process(StreamEvent event) {
    System.out.println("Input value: " + new String(Bytes.toBytes(event.getBody())));
}
```

CREATE LAUNCH CONFIGURATION

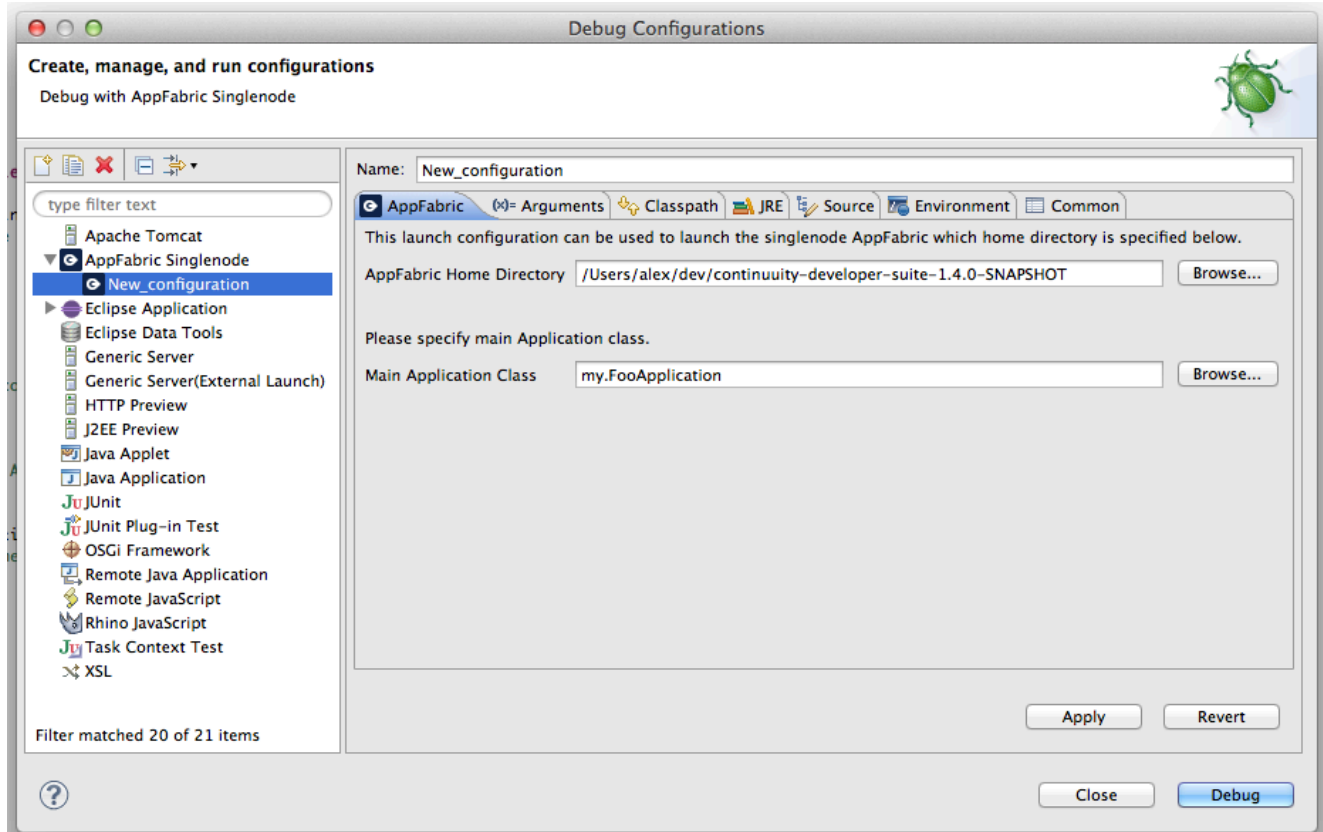
Right-click on “FooApplication” in the Project Explorer (usually the left panel), choose Debug As -> Debug Configurations from the popup menu.



Create a configuration for the "AppFabric Singlenode" top-level entry by right clicking on it or by clicking the “Add” icon above the configurations list.

Optionally, set the Name of the launch configuration (e.g. "FooApplication"), since it will be used to run this specific project.

On the first tab of the configuration, set the AppFabric Home Directory to the developer’s suite home and select the Main Application class (e.g. "my.FooApplication"). The class selection dialog autocomplete feature will help you once you start typing the class name.



Click Apply to save configuration.

START DEBUGGING

To start a debug session click Debug and you will see the output of the launched AppFabric in the Eclipse console. Wait for AppFabric to start and for the application to be deployed. Note in Eclipse's output console the application indicates as "successfully deployed", the last message in console should show "Finished deploy, exit code: 0".

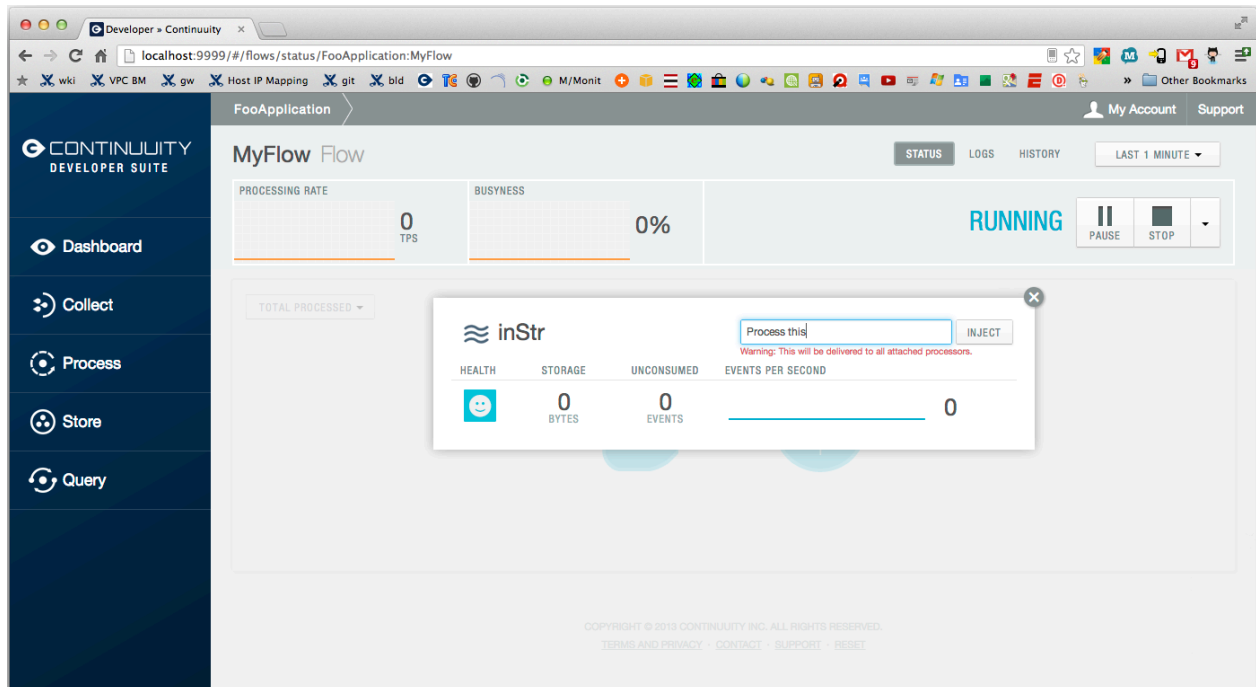
```

Console  Tasks  Search
New_configuration [AppFabric Singlenode] /System/Library/Java/JavaVirtualMachines/1.6.0.jdk/Contents/Home/bin/java (Feb 26, 2013 4:08:35 AM)
=====
Continuity AppFabric (tm) - Copyright 2012-2013 Continuity, Inc. All Rights Reserved.
=====
Continuity AppFabric started successfully
Connect to dashboard at http://ALEXMAC.local:9999
Deploying application...
OUTPUT > Deploying... ./Users/alex/dev/eclipse-workspace/FooApplication/FooApplication.jar
OUTPUT > Deployed
Finished deploy, exit code: 0

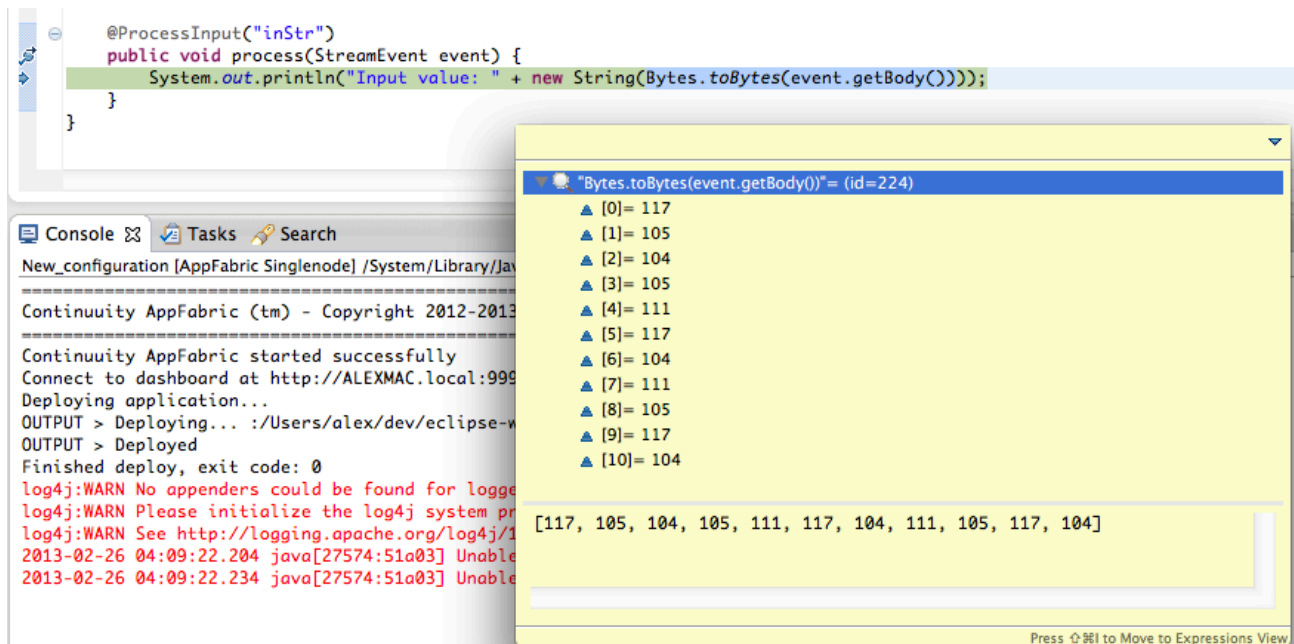
```

WORK WITH APPLICATION

Go to "<http://localhost:9999>" in your web browser. Navigate to "Process" using the menu on the left and start the Flow. To inject data into the Stream, click on "Text", enter some text and click "INJECT".



Eclipse will "catch" a breakpoint in `MyFlowlet.process()` method.



STOP DEBUG

The Debug session can be stopped in a standard way, e.g. clicking the Terminate button in the top menu icons panel. This will stop AppFabric process.

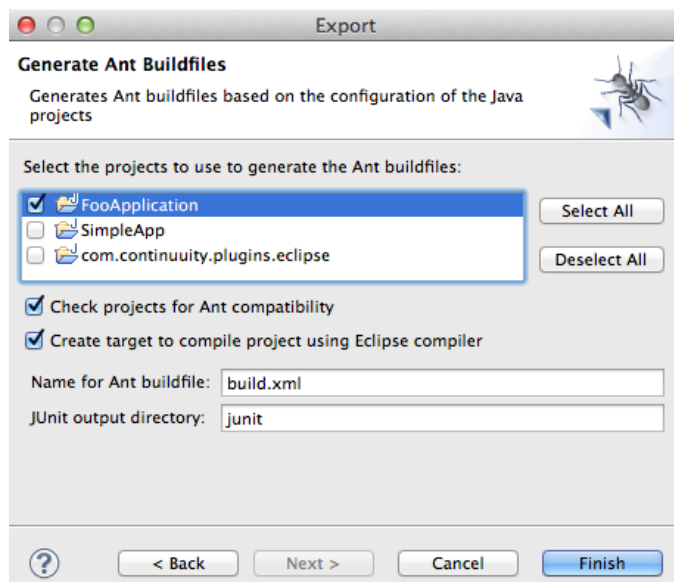
RUN MODE

The steps for running an application are the same as for debugging it. Select the “Run As” menu item instead of “Debug As”.

BUILDING A DEPLOYABLE JAR USING ANT

From Eclipse:

- Highlight project
- Right click select "Export..."
- Under "General" select "Ant Buildfiles" and click "Next"



- Select Checkbox for your project
- Check "Create target to compile project using Eclipse compiler"
- Click Finish
- Open "Build.xml" which should have been added to your project
- Before `</Project>` add the following:

```
<manifest file="MANIFEST.MF">
  <attribute name="Manifest-Version" value="1.0" />
  <attribute name="Main-Class" value="my.FooApplication" />
</manifest>
<jar destfile="FooApplication.jar" manifest="MANIFEST.MF">
  <fileset dir="bin">
    <include name="**"/>
  </fileset>
</jar>
```

- Replace `FooApplication` with the appropriate application class and package name.

7. CONCLUSION

Thanks again for downloading the Continuuity Developer Suite. By now you should be well on your way to building your Big Data application.

Once you have built your application, be sure to fire up your Developer Sandbox so you can push it to the cloud. To get your Developer Sandbox set up just go to <https://accounts.continuuity.com/>.

If you need any help from us along the way or have specific questions, please go to <http://support.continuuity.com>.