Data integration and analytics as a service, Part 1: DataWorks

Ahmed Abdel Monsef (amonisf@eg.ibm.com)
Advisory IT Specialist, IBM Certified Experienced IT Specialist, Technical Leader
IBM

10 July 2015

Most data integration specialists find that data loading and migration from a source to target are usually time-consuming and tedious tasks to perform. Now with the IBM Bluemix™ DataWorks service, you can load and migrate data from different sources to different targets easily. This tutorial shows you how.

Overview

In this tutorial, learn how to use the DataWorks cloud service in IBM Bluemix to perform data load or migration from different sources to various targets. IBM DataWorks service, which includes DataWorks APIs and DataWorks Forge, allows developers to load, cleanse and profile data, in addition to migrating to different targets seamlessly. DataWorks Forge is primarily for knowledge workers and helps them to select data, visualize, and prepare it for use after enriching and improving its quality. This tutorial is Part 1 of a series covering data integration and analytics as a service.


DataWorks includes DataWorks APIs, which are REST-based APIs used by developers to quickly access different data sources. The DataWorks APIs enable you to perform three main functions: data load, U.S. address cleansing, and data profiling. The available data sources/targets are listed in Table 1.

Supported databases

Table 1. Supported source and target database by IBM DataWorks APIs

<table>
<thead>
<tr>
<th>Sources</th>
<th>Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Redshift</td>
<td></td>
</tr>
<tr>
<td>Amazon S3 (CSV files)</td>
<td></td>
</tr>
<tr>
<td>Analytics for Hadoop (CSV files)</td>
<td></td>
</tr>
</tbody>
</table>
The IBM DataWorks service also includes DataWorks Forge, which makes all users, including business analysts, and non-technical users, capable of selecting required data sets and simply visualizing it to start analysis by automatically loading, profiling, and classifying data to improve its quality. DataWorks Forge provides solutions for removing duplicates, and filtering and joining different data sets. After that, you are ready to load the improved data sets to different set of targets. The available data sources/targets are listed below.

**Table 2. Supported source and target database by IBM DataWorks Forge**

<table>
<thead>
<tr>
<th>Sources</th>
<th>Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Redshift</td>
<td>dashDB</td>
</tr>
<tr>
<td>dashDB</td>
<td>IBM DB2</td>
</tr>
<tr>
<td>IBM DB2</td>
<td>IBM Informix</td>
</tr>
<tr>
<td>IBM Informix</td>
<td>MySQL</td>
</tr>
<tr>
<td>MySQL</td>
<td>Oracle</td>
</tr>
<tr>
<td>Oracle</td>
<td>Pivotal Greenplum</td>
</tr>
<tr>
<td>Pivotal Greenplum</td>
<td>PostgreSQL</td>
</tr>
<tr>
<td>PostgreSQL</td>
<td>Salesforce.com</td>
</tr>
<tr>
<td>Salesforce.com</td>
<td>SQL Database</td>
</tr>
<tr>
<td>SQL Database</td>
<td>Sybase</td>
</tr>
<tr>
<td>Sybase</td>
<td>Sybase IQ</td>
</tr>
</tbody>
</table>
As you think through the importance of enriching and improving the quality of the source data, you'll realize that many business entities, (not just technical teams) might be involved in improving data quality. So let us go through the process of how to accomplish this using the DataWorks service. Also be reminded that data shaping is also of prime importance, while improving the data quality, which can be easily applied using DataWorks service.

Let's use the data load, cleansing and profiling APIs using the below DataWorks sample applications, using Java runtime for Platform as a Service (PaaS) applications. The application uses DataWorks and any selected source database services, and the service-composition programming model supported by PaaS environments like Cloud Foundry. Let's use these sample applications at IBM Bluemix DevOps Services. This provides us with a place to publish the application.

### Table 3. Sample IBM Bluemix DevOps Services applications

<table>
<thead>
<tr>
<th>Sample applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Load sample application</td>
</tr>
<tr>
<td>Address Cleansing sample application</td>
</tr>
<tr>
<td>Data Profiling sample application</td>
</tr>
<tr>
<td>IBM Bluemix catalog</td>
</tr>
</tbody>
</table>

#### What you need

You will need the following to run the above listed sample applications:

1. An IBM Bluemix account
2. An IBM Bluemix DevOps Services account
3. A basic familiarity with databases

#### Got a Bluemix question?

Ask on Stack Overflow
Ask on dW Answers
What's the difference between asking on Stack Overflow and asking on dW Answers?

#### Running the Data Load API applications

The steps below describe the most basic approach to run DataLoad API sample applications (also see Detailed explanation for DataWorks APIs):

1. Step 1. Deploy the data-load sample application from IBM Bluemix DevOps Services.
   1.1. Fork dwaas-load application repository on your DevOps account: Deploying the application.
   1.2. Liberty for Java application will be added to your Bluemix dashboard with the same name of your cloned sample application.
Figure 1. Confirming that dwaas-load application is built and deployed successfully

Figure 2. Confirming that dwaas-load application is added to your Bluemix apps

2. Step 2. Add a new DataWorks service and bind it to the dwaas-load application.
   2.1. Add DataWorks service or API as shown below or by adding it separately from your dashboard and binding it to your application using Cloud Foundry, as shown in Listing 1.
   2.2. Restage dwaas-load application, as it is mandatory to be able to use new DataWorks service from your application.
Figure 3. Adding DataWorks to dwaas-load application

Figure 4. Restaging dwaas-load application

⚠️ Restage Application

Your ‘DWaaS’ app must be restaged to use the new ‘DWaaS’ service. Restaging makes this service available for use. Do you want to restage it now?

RESTAGE  CANCEL
Figure 5. Confirming that DataWorks service is bound to the dwaas-load application

Listing 1. Binding DataWorks service to dwaas-load application

cf api https://api.ng.bluemix.net
cf login -u user -o organization -s dev
cf bind-service dwaas-load DWaaS

3. Step 3. Add two SQL Database services and bind them to dwaas-load application.
You can use Cloud Foundry to bind the SQL Database service, as shown in Listing 2.

3.1. Add SQL Database services
3.2. Bind SQL Database services
3.3. Restage dwaas-load application, as it is mandatory to be able to use new added services from your application.

Figure 6. Binding SQL Database service to dwaas-load application
Figure 7. Restaging dwaas-load application

⚠️ Restage Application

Your 'dwaas-load' app must be restaged to use the new 'DWaaS-SQL-DB' service. Restaging makes this service available for use. Do you want to restage it now?

Listing 2. Binding SQL Database service to dwaas-load application

```bash
cf api https://api.ng.bluemix.net
cf login -u user -o organization -s dev
cf bind-service dwaas-load DWaaS-SQL-DB
```

   4.1. Get the SQL Database information and credentials from the Environment Variables VCAPSERVICES as predefined on your dwaas-load application. These are added once you bind your SQL Database to the application.
   4.2. Run and test the dwaas-load Liberty for Java application using the URL provided in your application.
Running the IBM DataWorks Address Cleansing sample application

The steps below describe the most basic approach to run a data-cleansing API sample application. For more advanced tooling to speed your development using those APIs, see Detailed explanation for IBM DataWorks APIs.

Deploying the IBM DataWorks Address Cleansing sample application

1. Step 1. Deploy the data-cleansing application.
   1.1. Fork dwaas-cleansing application repository on your DevOps account: Deploying the application from IBM Bluemix DevOps Services.
   1.2. Liberty for Java application will be added into your Bluemix dashboard with the same name of your cloned sample application.
2. Step 2. Add a new DataWorks service and bind it to the dwaas-cleansing application.
   2.1. Add DataWorks service or API as shown below or by adding it separately from your
dashboard, then bind it to your application, using Cloud Foundry, as shown in Listing 3.
   2.2. Restage dwaas-cleansing application, as it is mandatory to be able to use DataWorks
from your application.
Figure 11. Adding DataWorks service to dwaas-cleansing application

Figure 12. Restaging dwaas-cleansing application

⚠️ Restage Application

Your 'dwaas-cleansing' app must be restaged to use the new "DWaaS" service. Restaging makes this service available for use. Do you want to restage it now?
Figure 13. Confirming that DataWorks service is bound to the dwaas-cleansing application

Listing 3. Binding DataWorks service to dwaas-cleansing application

```bash
# Step 1
cf api https://api.ng.bluemix.net
# Step 2
cf login -u user -o organization -s dev
# Step 3
cf bind-service dwaas-cleansing DWaaS
```

Run and test the dwaas-cleansing Liberty for your Java application using the URL provided in your application.
Running the Data Profiling API sample application

The steps below describe the most basic approach to run the Data Profiling API sample application. For more advanced tooling to speed your development using those APIs, see Detailed explanation for IBM DataWorks APIs.

Deploying the Data Profiling API sample application

1. Step 1. Deploy the data-profiling application from IBM Bluemix DevOps Services.
   1.1. Fork the dwaas-profiling application repository on your DevOps account: Deploying the application.
   1.2. Liberty for Java application will be added to your Bluemix dashboard with the same name of your cloned sample application.
Figure 15. Confirming that dwaas-profiling application built and deployed successfully

Figure 16. Confirming that dwaas-profiling application is added to your Bluemix apps

2. Step 2. Add DataWorks and SQL Database services and bind them to the dwaas-profiling application.
   2.1. Add and bind DataWorks service or API and SQL Database service, as shown below or by adding it separately from your dashboard, then bind it to your application, using Cloud Foundry, as shown in Listing 4.
   2.2. Restage dwaas-profiling application, as it is mandatory to be able to use new DataWorks service from your application.
Figure 17. Adding DataWorks and SQL Database services to dwaas-profiling application

Figure 18. Restaging dwaas-profiling application

⚠️ Restage Application

Your 'dwaas-profiling' app must be restaged to use the new 'DWaaS' service. Restaging makes this service available for use. Do you want to restage it now?

RESTAGE CANCEL

Figure 19. Confirming that DataWorks and SQL Database services are bound to the dwaas-profiling application

Listing 4. Binding DataWorks and SQL Database services to the dwaas-profiling application

```bash
cf api https://api.ng.bluemix.net
cf login -u user -o organization -s dev
cf bind-service dwaas-profiling DWaaS
cf bind-service dwaas-profiling DWaaS-SQL-DB
```

Run and test the dwaas-profiling Liberty for Java application using the URL provided in your application.

**Figure 20. Run the application**

<table>
<thead>
<tr>
<th>Column</th>
<th>Column Name</th>
<th>Status</th>
<th>Cardinality</th>
<th>Internal Data Type</th>
<th>Format</th>
<th>Classification</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PRODMS</td>
<td>Completed</td>
<td>23</td>
<td>varchar</td>
<td>AAD000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ACCTNO</td>
<td>Completed</td>
<td>80</td>
<td>integer</td>
<td>00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ADDRESS</td>
<td>Completed</td>
<td>25</td>
<td>varchar</td>
<td>00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>POSTDATE</td>
<td>Completed</td>
<td>66</td>
<td>date</td>
<td>NULL/DEFAULT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ACCTNAME</td>
<td>Completed</td>
<td>80</td>
<td>varchar</td>
<td>00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Running the Data Forge service**

The steps below describe the most basic approach to run Data Forge Service. For more advanced tooling to speed your development using this service, see [Detailed explanation for Data Forge](#).

1. Step 1. Add the DataWorks service in order to use its Data Forge capability.

**Figure 21. Adding independent DataWorks service without any application**

![Add a Service or API](image)

**Figure 22. Confirming that DataWorks service is added to your Bluemix dashboard**

![DataWorks](image)

2. Step 2. Add Data Sources, which are subject for shaping.
3. Step 3. Examine the source data quality and discover which areas have to be improved. Then select the appropriate action to improve the quality of that source data.
Figure 27. Examine the quality of selected data

![Figure 27](image1)

Figure 28. Select appropriate action to increase the data quality

![Figure 28](image2)

Figure 29. Data quality increased from 70 percent to 87 percent after removing duplicates

![Figure 29](image3)
4. Step 4. Save shaped data to a target database and monitor your activity.

**Figure 30. Selecting target data source**

**IBM DataWorks Forge**

<table>
<thead>
<tr>
<th>Work with data</th>
<th>My activities</th>
</tr>
</thead>
</table>

Back to search

Select a data source to copy data to

- Cloudant
- dashDB
- IBM Watson Analytics
- SQL Database

**Figure 31. Monitoring your loading activity**

**IBM DataWorks Forge**

<table>
<thead>
<tr>
<th>Work with data</th>
<th>My activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activities</td>
<td>Monitoring</td>
</tr>
</tbody>
</table>

1 - 4 of 4

**Activity name**

| Remove_Duplicates |

**Conclusion**

We have seen how easy it is to work with capabilities like data load, data cleansing, profiling, and shaping using IBM DataWorks APIs and DataWorks Forge. These services provide easy access and control over multiple data integration processes. It's obvious why these services decrease the time consumed compared to the regular ways.

Stay tuned for Part 2, covering other Analytics services on Bluemix.
Resources

- The Information Management area on developerWorks provides resources for architects, developers, and engineers.
- Stay current with developer technical events and webcasts focused on a variety of IBM products and IT industry topics.
- Follow developerWorks on Twitter.
- Watch developerWorks demos ranging from product installation and setup demos for beginners, to advanced functionality for experienced developers.
- Get involved in the developerWorks Community. Connect with other developerWorks users while you explore developer-driven blogs, forums, groups, and wikis.
About the author

Ahmed Abdel Monsef

Ahmed Abdel Monsef is a data specialist, architect and technical lead and is part of the Business Analytics and Strategy team. He has been involved in the delivery of data migration and database administration projects for the past 10 years and has been responsible for a number of architecture projects and decisions in different industries, currently focusing on big data, analytics, and cloud computing technologies. @AMonsef82

© Copyright IBM Corporation 2015

Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)