

Move and Prepare Data With Ease Using IBM DataWorks Forge

1. Introduction

Setting up an enterprise level data movement infrastructure and then ingesting that data from a secure environment to a cloud repository can be challenging. As a Data Engineer, you need a simple point and click solution to move, discover, cleanse, standardize and transform data to support analytic applications. IBM Cloud Solutions with its DataWorks as-a-Service solves this challenge by seamlessly connecting, transforming and ingesting the data into the secure IBM Cloud.

The extremely intuitive and efficient DataWorks service, with its wide collection of data stores connections, prepares and ingests the data to the IBM cloud with ease and robustness.

In this lab we will explore how a data engineer utilizes DataWorks and other IBM Bluemix cloud services to easily host, connect, ingest and persist enterprise data on to the cloud.

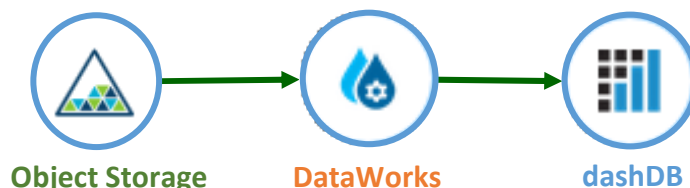
2. Lab Components: IBM Bluemix Cloud Offering Used

1. Source Data Repository: **Object Storage Service**

NOTE: This usually is the data hosted in a secure enterprise environment that could be securely retrieved through a network tunnel using the BlueMix Secure gateway. Since getting access to a data hosted in an enterprise infrastructure can be a compliance and privacy issue for this lab, we will be using sample datasets hosted in the IBM Object Storage that would be used as a source data set.

2. Transform and Migrate: **DataWorks Service**

3. Target Data Repository: **dashDB Service**



3. Before You Begin

1. Download Lab-DataWorks.zip archive from the github.com location below and extract the two data files (GOProductDim, GOSalesFact) to your laptop:


<https://github.com/ibmdataworks/sigmo-demos/tree/master/data-engineer/ingest>

2. Log in to your IBM Cloud Bluemix account
<https://console.ng.bluemix.net/login?state=/home/onboard/>

NOTE: If you don't have a Bluemix account then get started for free by registering using the URL <http://www.ibm.com/cloud-computing/bluemix/> or <https://console.ng.bluemix.net/registration/>

Start of Lab

4. Initialize Source Data Repository: Object Storage Service


1. From the Bluemix dashboard catalog menu search for "Object Storage"
2. Click on the Object Storage Icon 
3. Choose the default pre-filled values in the fields (optionally rename the "Service name:" to DE-DataIngest-ObjStorage), select the "Free Pricing Plan" and click "Create" at the bottom of the page.
4. Click on "Actions->Add Container" and enter "datafirst" as the container name.

NOTE: Once a new container is created, you get this message "There are no files in the container you selected. Add files or view Object Storage documentation."

5. Click on "Add files".


6. Select and add the files previously downloaded (GOProductDim, GOSalesFact) to this Container.

NOTE: Your data files are now moved to the Object Storage into the Bluemix.

7. From the Object Storage Service menu (scroll up the browser page), click on “Service Credentials” option.
8. Select the Key Name for your service (e.g. Credentials-1) and select the “View Credentials” down triangle () under the ACTIONS column.

It will show the credential details similar to the following:

```
{
  "auth_url": "https://identity.open.softlayer.com",
  "project": "object_storage_abcd12e3_4f56_7a89_1bc2_3de45fa67bcd",
  "projectId": "ab12c3456d7891ef2345a6789bcd1d23e",
  "region": "dallas",
  "userId": "123a4567890123456b789012cde34f56",
  "username": "admin_123a4567890123455a67ce456b789012cde34f56",
  "password": "ANUP&3NaZiR4PD&r=",
  "domainId": "cc421938a87348739a6bdc374424273d",
  "domainName": "1531219",
  "role": "admin"
}
```

9. Copy the credentials by clicking on the blue square () on the top right corner of the window and paste it into a notepad (Windows) / textedit (Mac) / vi (Linux). This information will be needed later for use in DataWorks to create and establish the connection to the object store.

From the credential details above, only the values from the following tokens (as highlighted in blue) will be needed in DataWorks to establish connection to the Object Store: "auth_url", "projectId", "region", "userId", "password"

10. The source data repository to be used with DataWorks later is now ready. Let’s create a target data repository.
11. From the top of the browser page click on the “Catalog” option. This will take you back to the Bluemix dashboard catalog menu.

5. Initialize Target Data Repository: dashDB Service

1. From the Bluemix dashboard catalog menu search for “dashdb”
2. Click on the dashDB Icon 

3. Choose the default pre-filled values in the fields (optionally rename the “Service name:” to DE-DataIngest-dashDB), select the “Entry Pricing Plan” (default) and click “Create” at the bottom of the page.
4. From the dashDB Service menu (scroll up the browser page), click on “Service Credentials” option (you may have to click on the double arrow button on the “Back to Dashboard for...” option).

It will show the credential details similar to the following:

```
{
  "credentials": {
    "port": 50000,
    "db": "BLUDB",
    "username": "dash999999",
    "ssljdbcurl": "jdbc:db2://awh-yp-small99.services.dal.ibmcloud.net:50001/BLUDB:sslConnection=true;",
    "host": "awh-yp-small99.services.dal.ibmcloud.net",
    "https_url": "https://awh-yp-small99.services.dal.ibmcloud.net:8443",
    "dsn": "DATABASE=BLUDB;HOSTNAME=awh-yp-small99.services.dal.ibmcloud.net;PORT=50000;PROTOCOL=TCPIP;UID=dash999999;PWD=ANupNAirAjxy1;",
    "hostname": "awh-yp-small99.services.dal.ibmcloud.net",
    "jdbcurl": "jdbc:db2://awh-yp-small99.services.dal.ibmcloud.net:50000/BLUDB",
    "ssldsn": "DATABASE=BLUDB;HOSTNAME=awh-yp-small99.services.dal.ibmcloud.net;PORT=50001;PROTOCOL=TCPIP;UID=dash999999;PWD=ANupNAirAjxy1;Security=SSL;",
    "uri": "db2://dash999999:ANupNAirAjxy1@awh-yp-small99.services.dal.ibmcloud.net:50000/BLUDB",
    "password": "ANupNAirAjxy1"
  }
}
```

5. Copy the credentials by clicking on the blue square (■) on the top right corner of the window and paste it into a notepad (Windows) / textedit (Mac) / vi (Linux). This information will be needed later for use in DataWorks to create and establish the connection to dashDB.

From the credential details above, only the values from the following tokens (as highlighted in blue) will be needed in DataWorks to establish connection to dashDB:
"host", "db", "username", "password"

NOTE: You can also retrieve the connection information by Opening the dashDB console (Manage->Open from dashboard) and within the console clicking on “Connect-> Connection information”. You get the information screen similar to the following:


Individual database settings:	
Host name:	awh-yp-small99.services.dal.ibmcloud.net
Port number:	50000
Database name:	BLUDB
User ID:	dash999999
Password:	ANupNAirAjxy1
Version:	Compatible with DB2 for Linux, UNIX, and Windows, Version 11.1 or later

6. The target data repository to be used with DataWorks later is now ready. Let's work with DataWorks to transform and migrate data.


NOTE: Although we have created the database service, it is optional to port the source schema to this target database. More information about this in the next section.

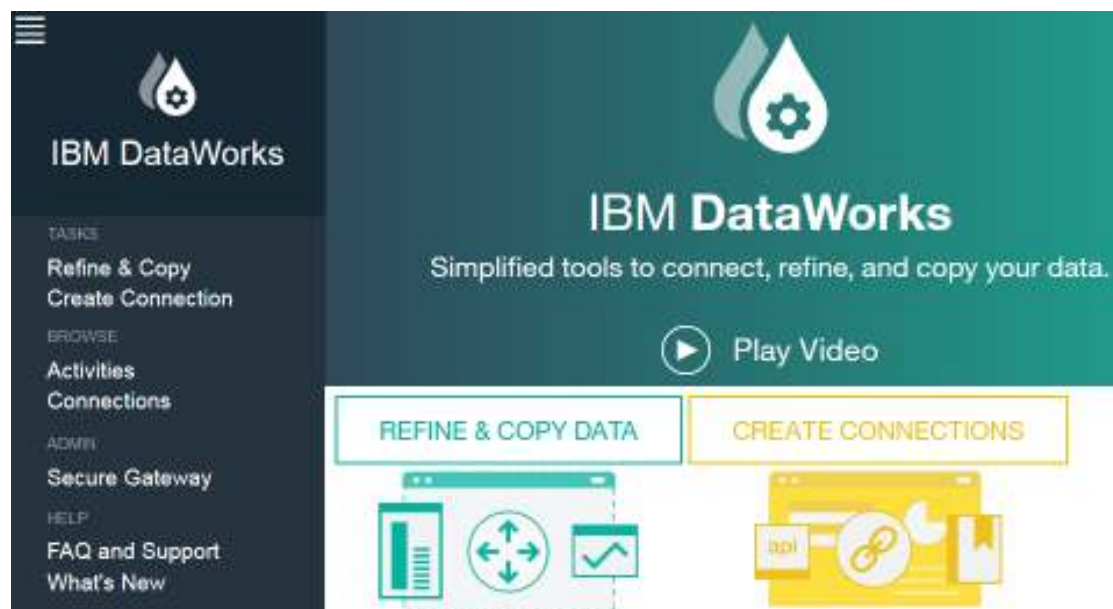
7. From the top of the browser page click on the "Catalog" option. This will take you back to the Bluemix dashboard catalog menu.

6. Transform and Migrate Data: DataWorks Service

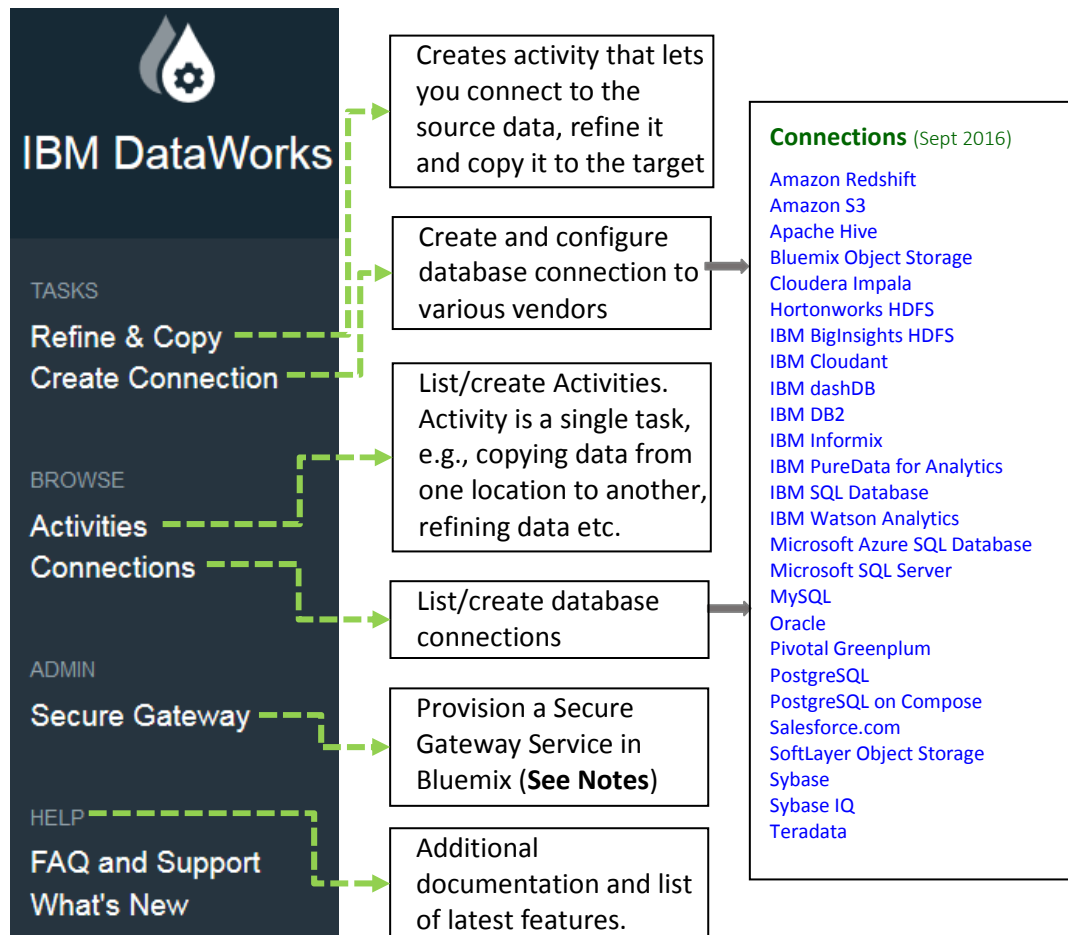
1. From the Bluemix dashboard catalog menu search for "dataworks"
2. Click on the DataWorks Icon 
3. Choose the default pre-filled values in the fields (optionally rename the "Service name:" to DE-DataIngest-DataWorks), select the "Starter Pricing Plan" (default) and click "Create" button on the side panel.

NOTE: Once the service is provisioned it will invoke the DataWorks Service dashboard where you can explore additional resources like IBM DataWorks™ APIs, Learning Center, Discussion forums etc.

4. To launch the DataWorks service just created, click the arrow icon () in the IBM DataWorks™ section on the provisioning dashboard. This brings you to the DataWorks dashboard which looks like this:

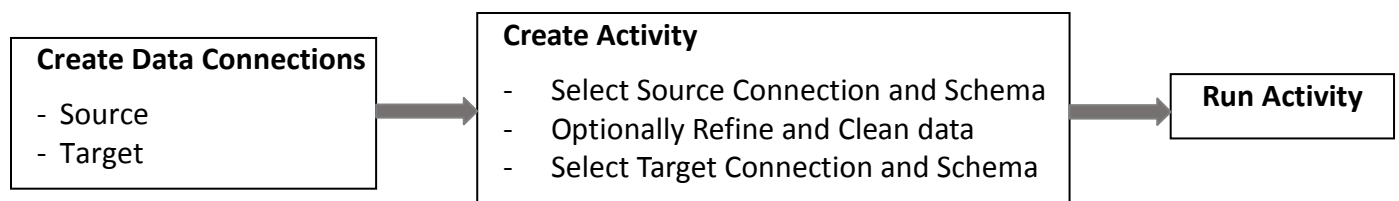


6.1 DataWorks Service Introduction



NOTE: The Bluemix Secure gateway can be used to retrieve data hosted in a secure enterprise environment through a secure network tunnel. Configuring a Secure Gateway service and connecting using a secure tunnel is a simple and intuitive process (the dashboard walks you through the steps). As it was indicated earlier, getting access to a data hosted in an enterprise infrastructure can be a compliance and privacy issue for this lab, we will be using a sample datasets hosted in the IBM Object Storage that would be used as a source data set. This doesn't need the use of a secure gateway.

The flow diagram for creating a data migration process in DataWorks is as shown below:




6.2 Create Data Connections

6.2.1 Create Source Data Connection

1. From the DataWorks dashboard click on “Tasks”->“Create Connection”

This will take you to the list of all available connection drivers.

2. Select the “Bluemix Object Storage” icon ()
3. Enter the following information in the data entry dashboard

Connection Name: SourceDB

Description: Data Stored in Bluemix Object Storage

4. Retrieve the values for the parameters "auth_url", "projectId", "region", "userId", "password" from the Bluemix Object Storage service that was created earlier (section titled “Initialize Source Data Repository: Object Storage Service” step 8) and enter it in the dashboard.

Your entry would look similar to the following:

Auth URL: `https://identity.open.softlayer.com`
Project ID: `ab12c3456d7891ef2345a6789bcd1d23e`
Region: `dallas`
User ID: `123a4567890123456b789012cde34f56`
Password: `ANUP&3NaZiR4PD&r`


5. Click on the “CREATE CONNECTION”
6. You will be seeing the **SourceDB** icon in the Connections dashboard.



6.2.2 Create Target Data Connection

1. From the DataWorks dashboard click on “Tasks”->“Create Connection”

This will take you to the list of all available connection drivers.

2. Select the “IBM dashDB” icon ()
3. Enter the following information in the data entry dashboard

Connection Name: TargetDB
Description: Data Stored in IBM dashDB

4. Retrieve the values for the parameters "host", "db", "username", "password" from the dashDB service that was created earlier (section titled "Initialize Target Data Repository: dashDB Service" step 4) and enter it in the dashboard.

You entry would look similar to the following:

Hostname or IP Address: awh-yp-small199.services.dal.bluemix.net
Database: BLUDB
Username: dash9999999
Password: ANupNAirAjxyl

5. Click on the "CREATE CONNECTION"
6. You will be seeing the **TargetDB** icon in the Connections dashboard.



6.3 Create Data Flow (Activity)

1. From the DataWorks dashboard click on "Tasks"->"Refine and Copy"
2. Click and select the source database: SourceDB
3. Click and select the Container: datafirst
4. Click and select the files (checkbox) from the container: GOProductDim, GOSalesFact

At this point you have the options of either **copying/migrating the selected data files as-is** to the target OR **refine the data** from the selected files and then copy/migrate it.

6.3.1 Copy data as-is

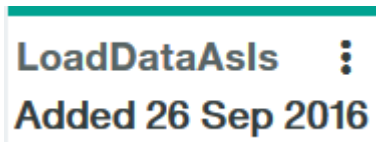
5. Edit the activity name by clicking the pencil icon (✎) on top of the dashboard. Name the activity "LoadDataAsIs". Click ✓ when done with editing.
6. Click on "COPY TO TARGET"
7. Click and select the target database: TargetDB
8. Select the database schema to which the data needs to be copied to.

NOTE: Since we are using an entry plan dashDB for this lab exercise the schema name would look similar to the dashDB userid (e.g., DASH99999)

9. In the “Table Action” column you are prompted with 4 options “Append to the table”, “Recreate the table”, “Replace the table contents” and “Merge with the table”. Since this is the first time we are copying data, select “Recreate the table”

NOTE: If the table does not exist, it will be created and data will be appended.

10. Now you can either “SCHEDULE” this activity for a later run or “RUN” this activity immediately. Click on “RUN” to execute it immediately
11. The activity is executed immediately. An icon is added to the activity list dashboard. We will revisit this icon in the later section.



IMPORTANT: You would have noticed that while creating the target database we never ported the schema. This is because, if the schema doesn't exist in the target database, it is created automatically from the source data type. This feature enhances and expedites the overall project productivity.

6.3.2 Refine Copy data as-is (OPTIONAL)

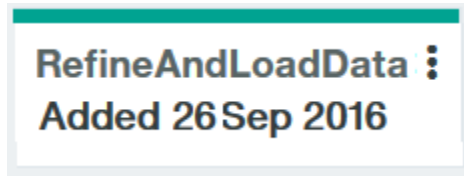
NOTE: Skip to next section titled “**Validate Data Flow Run Activity**” if you don't want to refine the data before copying else continue.

12. Repeat steps 1 to 4. **Come back to this step once complete.**
13. Edit the activity name by clicking the pencil icon (✎) on top of the dashboard. Name the activity “RefineAndLoadData”. Click ✓ when done with editing.
14. Click on “REFINE DATA”
15. This brings you to the refine dashboard where you can do various operations on the selected table.

For the tables, it provides you a histogram and statistics of the data quality for each of the table columns. You can cleanse the data (standardize values, remove rows/columns, change value, filter and sort rows etc.)

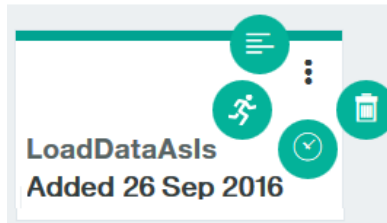
You are at liberty to explore and act on these options. Once complete click on “NEXT”

16. Continue steps 7 to 10. **Come back to this step once complete.**
17. The activity is executed immediately and an icon is added to the activity list dashboard.



6.4 Validate Data Flow Run Activity

1. From the DataWorks dashboard click on “Browse”->“Activities”
2. Click on the three dots (⋮) on the right-top corner for the activity titled “LoadDataAsIs”. This will pop up additional sub-activity icons



You can Run (🏃), Schedule (🕒), delete (🗑️) the activity OR find more details (⋮) about the run.

3. Let find out about the detailed status of the run. Either click on the activity name icon “LoadDataAsIs” or click on the expanded details (⋮) icon.

LoadDataAsIs

Activity Created: 26 September 2016

Description: [Edit]

Schedule: No schedule set. **SCHEDULE**

Activity Runs: All runs: 1 **RUN ACTIVITY**

Activity Input:

- GOPProductDev.exe SourceDB/datafirst
- GOSalesPart.exe SourceDB/datafirst

Activity Output:

Selected Target: TargetDB/DM2H6352

Action replace was successful.

GOPRODUCTDM_CSV View data

GOSALESFACT_CSV View data

Creator: anupr@paulm.com

Execution Log:

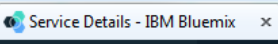

Message	Time
Enter spark-submit.sh	26 September 2016 23:...
Submission ID: driver-0710017213152-3509-6737b02...	26 September 2016 23:...
SparkRunner: job execution started	26 September 2016 23:...
IBM DataWorks Spark Engine version: 2.0.0.02	26 September 2016 23:...
SparkContext: Running Spark version 1.5.0	26 September 2016 23:...
SparkRunner: job execution completed in 13425 ms	26 September 2016 23:...
Exit spark-submit.sh	26 September 2016 23:...
Driver Status is FINISHED	26 September 2016 23:...
SparkRunner: job execution ended	26 September 2016 23:...

It will show you an information screen (similar to the figure above). Here, check if the activity details match with the information it was created with and whether the activity was run successfully. You can also rerun or schedule a run of the activity

without having to recreate it again. You can see the detailed **runlog** of the activity by clicking on the named “Activity Runs” (circled in the earlier figure).

6.5 Validate Migrated Data

1. Go to the main Bluemix portal and click on DASHBOARD (Top of the browser)

NOTE: You may already have a browser tab open from where you invoked the DataWorks service (). Go to that tab and click on . This will take you back to the provisioned services dashboard.

2. Under the Services section on this browser page click on the dashDB service that was created in section titled “Initialize Target Data Repository: dashDB Service “.
3. This will open up the dashDB launch board. Click on the “OPEN” button. This will launch the dashDB dashboard.
4. To verify the data you have 2 Options

Option 1

- a. Click on “Run SQL” and execute the “SELECT COUNT(*) FROM *<SchemaName.TableName>*” SQL commands against the GOPRODUCTDIM and GOSALESFACT table name. The resultant row counts should be 274 and 10000 rows respectively.
- b. You can also run other SELECT commands to check if the resultant data is valid especially if you have refined the incoming data.

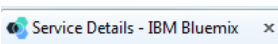

NOTE: For this lab the *SchemaName would be the* username as indicated in section titled Initialize Target Data Repository: dashDB Service step 4.

Option 2

- a. Click on “Tables”
- b. Select the Schema
- c. For each table GOPRODUCTDIM and GOSALESFACT verify the schema and also browse the data.

6.6 Delete All Provisioned Services (Optional)

1. Go to the main Bluemix portal and click on DASHBOARD (Top of the browser)

NOTE: You may already have a browser tab open from where you invoked the dashDB service (). Go to that tab and click on . This will take you back to the provisioned services dashboard.

2. Under the Services section you will see the three services you had created (Object Storage, DataWorks and dashDB)
3. To delete the service click on the three dots (⋮) under the actions column and click “Delete Service”. This will pop up a confirmation dialog box that says:

Deleting the service will remove it from all apps that are using it. In addition, all of its data will be permanently deleted. Are you sure you want to delete the <service name> service?
4. Click “Delete” to remove the service.
5. Repeat steps 3 and 4 above until all services are deleted.

End of Lab
