

Features of IBM DataWorks Forge in IBM Bluemix



After you complete this unit, you should understand:

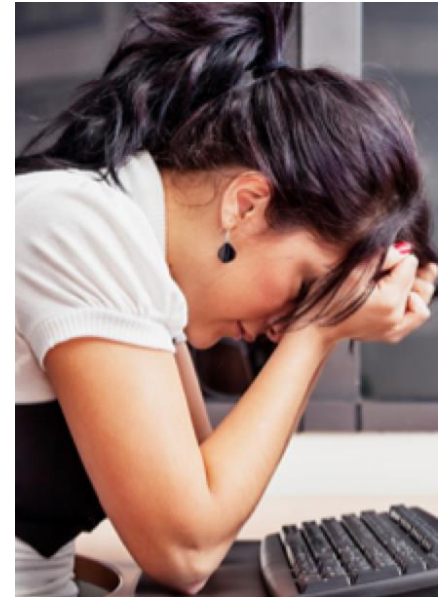
- Key characteristics of IBM DataWorks Forge Service in Bluemix
- How DataWorks Forge improves data collection and analysis

IBM DataWorks Forge: the IBM data refinery

- Provides a lightweight, cloud-based set of data refinement and access services
 - Makes fit-for-purpose data quickly and easily available to everyone across the enterprise.
- Ingests raw data, and provides services to enable you to perform:
 - Data preparation
 - Data movement and delivery
 - While also meeting associated privacy and security concerns before distributing refined data to end users.

Why Forge? How does Carol, a business analyst, go about analyzing data?

- Must ask IT for data and wait for it
- Ends up with raw data of undetermined quality
- If the data is the wrong data or the data sets are incomplete, she must wrangle with IT (more waiting)
- Culls through and joins data
- Gets the prepared data sets into a database or into analytics applications such as IBM Watson Analytics



IBM DataWorks Forge revolutionizes the experience for knowledge workers

**The better way:
IBM DataWorks Forge**



- Carol is empowered to quickly find relevant data herself.
- Provides automatically profiled and classified data
- Delivers quality scores and value distributions so that Carol can visualize and understand the data
- Allows Carol to quickly and easily shape, enrich, and improve quality by joining, standardizing, filtering, removing duplicates, and so on
- Allows Carol to apply her changes and deliver the data to her chosen target or targets

IBM DataWorks Forge supported sources and targets

Sources	Targets
dashDB	Cloudant NoSQL DB
IBM DB2	dashDB
Oracle	IBM Watson Analytics
Salesforce.com	SQL Database
SQL Database	
Cloudera Impala	
Amazon Redshift	
IBM Informix	
IBM Netezza	
Microsoft Azure	
Microsoft SQL Server	
MySQL	
Oracle	
Pivotal Greenplum	
PostgreSQL	
Sybase	
Sybase IQ	

IBM DataWorks Forge at work

IBM DataWorks Forge

Work with data | My activities

Classification

Quality Score

7 Columns
CUSTOMERS 7 Columns

7

Column Type Categories

High Data Quality
4 columns contain non-standard missing values.
3 columns contain suspect values.
3 more reasons...

83

Medium Data Sample Size
1,000 Records
7 Columns

Medium

CUSTOMERS

Undo Cancel Save and Continue

+	ADDRESS1	CITY	COUNTRY	CUSTNAME	CUST_ID	POSTAL_CODE	STATE
	2057 Hannah Street	Bunnaloo			10,474	2,731	NSW
	981 Ferguson Street	Burbank			10,475	91,502	CA
	2484 Robinson Lane	Burbank			10,476	91,505	CA
	595 Chicago Avenue	Burbank			10,477	91,502	CA
	2941 Wood Street	Burkeville			10,478	23,922	VA
	4775 Southern Street	Burleson			10,479	76,028	TX
	443 Nutter Street	Burlingame			10,480	94,010	CA
	465 Lynn Ogden Lane	Burlington			10,481	1,803	MA
	4122 Eastland Avenue	Burlong	AU	Machelle Watkin	10,482	6,401	WA
	3239 Rosemont Avenue	Burnaby	CA	Claudia Higgins	10,483		BC
	1756 Apple Lane	BURNHOUSE	UK	Mattie Marsh	10,484		
	400 Avenue of the Stars	Burnside	US	Michael Flanagan	10,485	61,057	IL

As the data is ingested into IBM DataWorks Forge, users gain valuable insights, such as quality scores, value distributions, and data classification/profiling

Improve data quality with IBM DataWorks Forge

IBM DataWorks Forge



Work with data My activities

14 Columns
CUSTOMERS 14 Columns

14

Column Type Categories

a # ✓ ✖
9 4 1 0



Medium Data Quality

9 columns contain missing values.
9 columns contain non-standard missing values.

[2 more reasons...](#)


Large Data Sample Size

1,039 Records
14 Columns

Large

CUSTOMERS

Undo Cancel Save and Continue

	ADDRESS1	CITY	COUNTRY_CODE	CUSTNAME	CUST_ID	POSTAL_CODE	STATE	FREIGHT_CHARGES	ORDER_DATE
	2057 Hannah Street	Bunnaloo	AU						
	981 Ferguson Street	Burbank	US						
	2484 Robinson Lane	Burbank	US						
	595 Chicago Avenue	Burbank	US						
	2941 Wood Street	Burkeville	US						
	4775 Southern Street	Burleson	US						
	443 Nutter Street	Burlingame	US						
	465 Lynn Ogden Lane	Burlington	US		10,481	1,803	MA		
	4122 Eastland Avenue	Burlong	AU	Machelle Watkin	10,482	6,401	WA		
	3239 Rosemont Avenue	Burnaby	CA	Claudia Higgins	10,483		BC	28.55	2/29/08 0:00
	1756 Apple Lane	BURNHOUSE	UK	Mattie Marsh	10,484				

- Sort
- Filter
- Remove duplicates
- Drop column

- Improve data quality by filtering, removing duplicates
- Remove unnecessary columns
- Visualize clusters of data using sort

Enrich and experiment with data with IBM DataWorks Forge

The screenshot displays the IBM DataWorks Forge interface. At the top, the title bar reads 'IBM DataWorks Forge'. Below it, a navigation bar shows 'Work with data' and 'My activities'. The main area features a data table titled 'CUSTOMERS' with columns: ADDRESS1, CITY, COUNTRY_CODE, and CUS. The table contains 13 columns and 361 records. A 'Review actions' dialog box is open, listing four actions: 03 Join data sets (CUSTOMERS, ORDERS), 02 Filter (by COUNTRY_CODE), 01 Sort (by COUNTRY_CODE, Ascending), and 00 Drop column (POSTAL_CODE from CUSTOMERS). A blue callout box points to the dialog with the text: 'Review the actions applied to the data and experiment by removing previous actions and adding new ones'. The background table shows data for various customers, including Jennifer, Heather, Elizabeth, Ruth, Deborah, Kathleen, and Janette.

IBM DataWorks Forge

Work with data My activities

13 Columns
CUSTOMERS 13 Columns

Column Type Categories
a # 1 0
9 3 1 0

Medium Data Quality

Medium Data Sample Size
361 Records
13 Columns

Medium

CUSTOMERS

ADDRESS1	CITY	COUNTRY_CODE	CUS
2592 Lynch Street	Manchester	US	Jennifer
3452 Twin House Lane	Manhattan	US	Heather
4258 Dark Hollow Road	Mankato	US	Elizabeth
4424 Goldleaf Lane	Mansfield	US	Ruth Ne
4397 Fleming Street	Mantorville	US	Deborah
726 University Hill Road	Mantzville	US	Kathleen
154 Davis Place	Marble Hill	US	Janette
552 Formula Lane	Marion	US	Larry Gravely
981 Ferguson Street	Burbank	US	Sandra Mattingly
2484 Robinson Lane	Burbank	US	Kathleen Bell
595 Chicago Avenue	Burbank	US	Pablo Hardy
2941 Wood Street	Burkeville	US	Christopher Wrenn
74 Duquesne Road	Marysville	US	Melissa Davis

Review actions

- 03 Join data sets
CUSTOMERS, ORDERS
- 02 Filter
by COUNTRY_CODE
- 01 Sort
by COUNTRY_CODE, Ascending
- 00 Drop column
POSTAL_CODE from CUSTOMERS

Close

Undo Cancel Save and Continue

Review the actions applied to the data and experiment by removing previous actions and adding new ones

Summary

- IBM DataWorks Forge service in Bluemix allows you to:
 - Identify relevant data
 - Transform the Data to suit your needs
 - Load it to a system for use