



Features of IBM DataWorks Forge in IBM Bluemix



After you complete this unit, you should understand:

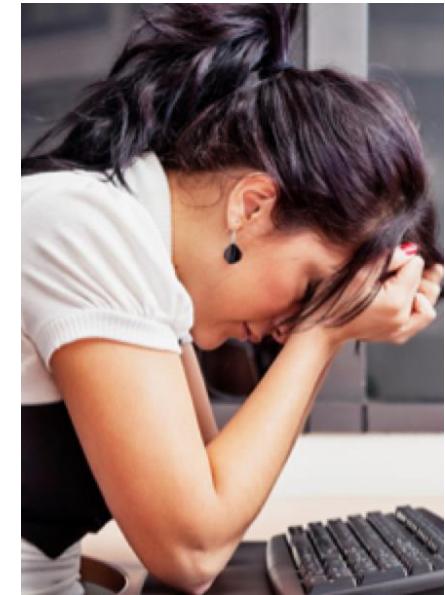
- Key characteristics of IBM DataWorks Forge Service in Bluemix
- How DataWorks Forge improves data collection and analysis

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

Classification Quality Score

Work with data My activities

7 Columns CUSTOMERS 7 Columns

Column Type Categories: **a** 5, **#** 2, **?** 0, ***** 0

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

Medium Data Sample Size: 1,000 Records, 7 Columns

Medium

CUSTOMERS

	ADDRESS1	CITY	STATE	ZIPCODE	CUSTNAME	CUST_ID	POSTAL_CODE	STATE
1	2057 Hannah Street	Bunnaloo				10,474	2,731	NSW
2	981 Ferguson Street	Burbank				10,475	91,502	CA
3	2484 Robinson Lane	Burbank				10,476	91,505	CA
4	595 Chicago Avenue	Burbank				10,477	91,502	CA
5	2941 Wood Street	Burkeville				10,478	23,922	VA
6	4775 Southern Street	Burleson				10,479	76,028	TX
7	443 Nutter Street	Burlingame				10,480	94,010	CA
8	465 Lynn Ogden Lane	Burlington				10,481	1,803	MA
9	4122 Eastland Avenue	Burlong	AU		Machelle Watkin	10,482	6,401	WA
10	3239 Rosemont Avenue	Burnaby	CA		Claudia Higgins	10,483		BC
11	1756 Apple Lane	BURNHOUSE	UK		Mattie Marsh	10,484		
12	400 Amherst Drive	Dove Ridge	US		Karen Fisher	10,485	61,257	IL

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

Undo Cancel Save and Continue

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, o 1, * 0

Medium Data Quality: 9 columns contain missing values. 9 columns contain non-standard missing values. 49

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		
1	2057 Hannah Street	Bunnaloo	AU								
2	981 Ferguson Street	Burbank	US								
3	2484 Robinson Lane	Burbank	US								
4	595 Chicago Avenue	Burbank	US								
5	2941 Wood Street	Burkeville	US								
6	4775 Southern Street	Burleson	US								
7	443 Nutter Street	Burlingame	US								
8	465 Lynn Ogden Lane	Burlington	US								
9	4122 Eastland Avenue	Burlong	AU	Machelle Watkin							
10	3239 Rosemont Avenue	Burnaby	CA	Claudia Higgins							
11	1756 Apple Lane	BURNHOUSE	UK	Mattie Marsh							
					10,481	1,803	MA				
					10,482	6,401	WA				
					10,483		BC	28.55	2/29/08 0:00		
					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

IBM DataWorks Forge

Work with data My activities

13 Columns CUSTOMERS 13 Columns 13 Column Type Categories a # ✓ ★ 9 3 1 0 Medium Data Quality 51 Missing values. Medium Data Sample Size 361 Records 13 Columns Medium

CUSTOMERS

ADDRESS1	CITY	COUNTRY_CODE	CUS
2592 Lynch Street	Manchester	US	Jennife
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	Debra
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 Business Park	Maynes	US	Malissa 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

FREIGHT_CHARGES ORDER_DATE ORDER_ID

38.99 11/11/08 0:00 8.958

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