

Modern Data Management with Boomi

Accelerates Business Outcomes

Abstract

This white paper provides overview, use case, integration, and architecture information for deploying Boomi Data Catalog and Preparation (DCP) 4.4 software on Dell EMC PowerEdge servers and Dell EMC PowerSwitch networking.

Data-Centric Workloads & Solutions

Notes, cautions, and warnings

 **NOTE:** A NOTE indicates important information that helps you make better use of your product.

 **CAUTION:** A CAUTION indicates either potential damage to hardware or loss of data and tells you how to avoid the problem.

 **WARNING:** A WARNING indicates a potential for property damage, personal injury, or death.

Chapter 1: Executive summary.....	5
Business challenge.....	5
Solution benefits.....	5
Intended audience.....	6
We value your feedback.....	6
Chapter 2: Use case.....	7
Overview.....	7
Starting point.....	7
Boomi capabilities used.....	8
Data sources and data stores.....	8
Data stores.....	8
Data sources.....	9
Data sets.....	10
Users and groups.....	10
Roles.....	11
Example users.....	12
Exploring data sets.....	13
Use case workloads.....	14
Start the use case sequence.....	15
Select a data set.....	15
Join data sets.....	16
Enrich the data.....	17
Choose columns.....	19
Add a filter.....	20
Aggregations.....	21
Output.....	22
Schedule a job.....	23
RESTful API example.....	24
Using Postman to access Boomi.....	24
Using a programming language to access Boomi.....	25
Chapter 3: Integrating Boomi DCP with Hadoop.....	27
Overview.....	27
Boomi server configuration.....	27
General specifications.....	27
Disk specifications.....	27
Software specifications.....	28
Hadoop specifications.....	28
Compatible Dell EMC hardware.....	29
Chapter 4: Solution architecture.....	30
Overview.....	30
Hadoop cluster.....	30

Boomi DCP server node functions.....	30
Networking.....	31
Metadata catalog.....	31
Annotation.....	31
Curation.....	32
Tagging.....	32
Chapter 5: Conclusions.....	33
Summary.....	33
Chapter 6: References.....	34
Dell Technologies documentation.....	34
Boomi documentation.....	34
Cloudera documentation.....	34
Dell Technologies InfoHub.....	35
More information.....	35

Executive summary

Data is quickly becoming the second most valuable resource for many organizations, after people resources. The importance that organizations place on cataloging and organizing the many available data assets into a single source of "truth" that is useful for analytics is evidence of the perceived value.

Topics:

- [Business challenge](#)
- [Solution benefits](#)
- [Intended audience](#)
- [We value your feedback](#)

Business challenge

Significant progress has been made in data consolidation that is associated with data lake initiatives. However, the ongoing creation of new systems hosting silos of information continues to complicate the efforts of both IT professionals and the data consumer communities. It is becoming increasingly clear that having a single, consolidated data analytics repository may not be a realistic goal. In order to enable a broad community of data consumers, organizations must provide tools that can handle the complexity of these complex data environments.

The proliferation of multiple data initiatives and new data silos appearing outside of any data lake boundary creates an environment where many potential data consumers cannot find or use critical information. Unfortunately, the systems that provide the best features for consolidating enterprise class data into data lakes are typically designed for professional data engineers who are comfortable with software development. The latest generation of tools available to transform, or "munge", data into consolidated lakes have greatly improved in scalability. However, they have not addressed ease-of-use limitations that prevent usage by many types of users with important, data-driven business use case needs.

In order to realize an attractive positive ROI from data consolidation investments by organizations, there must be a parallel effort to enable efficient data access for everyone with a demonstrable business need. This effort requires a differentiated set of tools for data discovery and access that must match the skills of diverse target users.

Most organizations today lack such an efficient process, where large groups of business users can uncover valuable data relationships between the information in a data lake and the many "satellite" data sources. The typical approach today relies too heavily on investments from IT in order to provide data extracts to feed the needs of people familiar with the relevant business concepts. Those users lack the required source system access and the coding skills to explore the data on their own.

Solution benefits

For too many organizations, the move to big data technologies has created barriers to information access for business and data analysts without formal software development training. This document shows how Boomi DCP can enable a broader user community to derive business value from practically any combination of enterprise data sources, without learning to program.

Boomi Data Catalog and Preparation (DCP) is a platform in which both IT and the business analyst community can find value. IT can easily add Boomi DCP to a new or existing big data cluster as easily as adding a traditional node to Hadoop. IT staff responsible for data governance can configure the right data access for the right users, using DCP role-based security integrated with existing source systems controls.

Business analysts using DCP can create powerful data transformations and summaries, with basic relational data structure and common SQL commands knowledge. These jobs are then run using highly scalable Hadoop services to produce new data artifacts. The artifacts are suitable for use with many popular reporting and visualization tools, both inside and outside of the Hadoop ecosystem.

Boomi DCP provides a platform that IT can easily install and manage, easy for data governance management, and easy for business analysts to use. This combination enables organizations that are investing in big data technologies to safely expand the end-user base, provide more value, and improve the ROI of any data-driven initiatives.

Intended audience


This document is intended for data center managers and IT architects who are involved with engineering, operation, or planning for Boomi Data Catalog and Preparation (DCP) on Dell EMC infrastructure.

This document assumes some familiarity with Boomi DCP capabilities and functions.

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on the solution and the solution documentation. Contact the Dell Technologies Solutions team by [email](#) or provide your comments by completing our [documentation survey](#).

Authors: Dell Technologies Data-Centric Workloads Engineering and Technical Marketing teams

 **NOTE:** For links to additional documentation for this solution, see the [Dell Technologies Solutions InfoHub for Data Analytics](#).

Use case

Dell Technologies has developed a complete end-to-end use case, that enables understanding of a typical Boomi DCP workflow.

Topics:

- [Overview](#)
- [Starting point](#)
- [Boomi capabilities used](#)
- [Data sources and data stores](#)
- [Users and groups](#)
- [Exploring data sets](#)
- [Use case workloads](#)
- [Schedule a job](#)
- [RESTful API example](#)

Overview

The complexity and amount of the data that Dell Technologies uses is not representative of anything that is encountered in an enterprise setting. The purpose of presenting this flow as a use case is to show the different types of problems that can be solved using only the Boomi DCP toolset.

This use case follows a fictitious consumer loan department that uses a combination of business systems. Each system requires a separate data source. Information that is related to customers (borrowers) and the loan details are stored in a PostgreSQL database with a simple one-table-per-object schema. Dell Technologies represents customer communications with a call center as a flat file extract stored on a file on a server.

The use case shows how the data discovery, data cataloging, and data transformation capabilities of Boomi DCP can be combined into an integrated business solution. The resulting workflow combines information from these source systems. It uses data query to join, filter, and transform that data into a result that can be used to generate a finished report. The final tabular result is written onto an output HDFS data store.

This example uses two structured data sources hosting multiple tables together with a file to create an output. The output lists how many customer communications there were for every approved and funded loan application. This use case demonstrates only a small set of the complete functionality available in Boomi DCP. This pattern is difficult to produce without the right toolset.

Starting point

This use case starts with multiple tables from a PostgreSQL database, and the contents from a file for the source data systems. The `stg_cust` source is our anchor table, with customer information including `custid`. The `stg_loan` table has information about loans including a `custid` foreign key into the customer table. The `stg_cc` file is in CSV format, and stored on a standard, nondistributed file system. The goal is to construct a list of how many customer communications existed for each successful loan. The output will include:

- The merged full name
- Total amount of loans
- Number of customer communications

Boomi capabilities used

Dell Technologies uses multiple capabilities in this use case workload. Data is pulled from multiple types of sources. The use case uses the integrated visualization of existing data, and the data preview window. A transformation is performed using Boomi's interface to a Hadoop cluster. Standard SQL transformations and outputs are also performed.

Data sources and data stores

Most large enterprise systems that are accessed through DCP have multiple collections of data objects. There must be a mapping between DCP and the data store to control outside access. The first level of defining a data connection is an adapter. Adapters are source-specific connectors that are developed for the product and distributed with the software.

An adapter must be enabled and configured before creating any higher-level data access objects. Boomi uses three additional object levels for data access control:

- Data stores
- Data sources
- Data sets

An adapter serves as a connecting code that enables DCP to communicate with the data store. To make a data store available to users, a Boomi administrator (or someone with equivalent privileges) uses the **Data Stores** page to add an external data location to the Boomi system. This location is a pointer to the location of the data and includes items like the machine name, and on what network port it is accessible. To access the pages listing the defined stores and sources in the system:

1. Click the **Tools > Manage** menu.
2. Select either **Data Stores** or **Data Sources**.

Data stores

Data stores are locations of data, such as:

- Databases
- Files located in various types of file systems
- Other such items

See [Data stores](#).

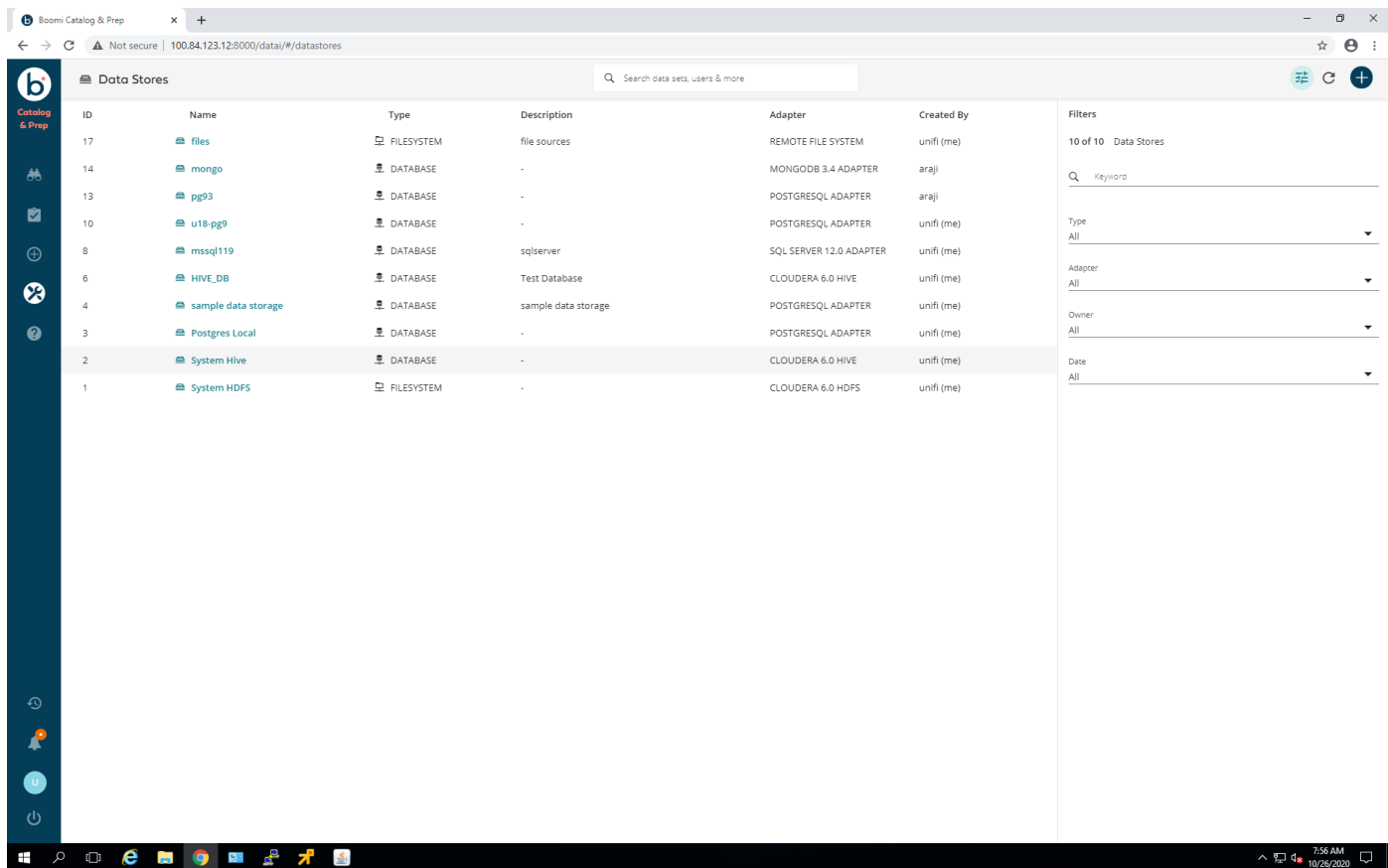


Figure 1. Data stores

Data sources

Data sources are definitions of a service, database, or file that is accessible by certain users. A definition includes any credentials that are required to access the data store, and which Boomi DCP users can see and use it. The definition is also where you establish whether the data source can also be a data sink, or writable repository. Having predefined users and groups with permissions before this step simplifies the creation of the data source. It is possible to define, on a per-column basis, who can see what portion of the data. See [Data sources](#).

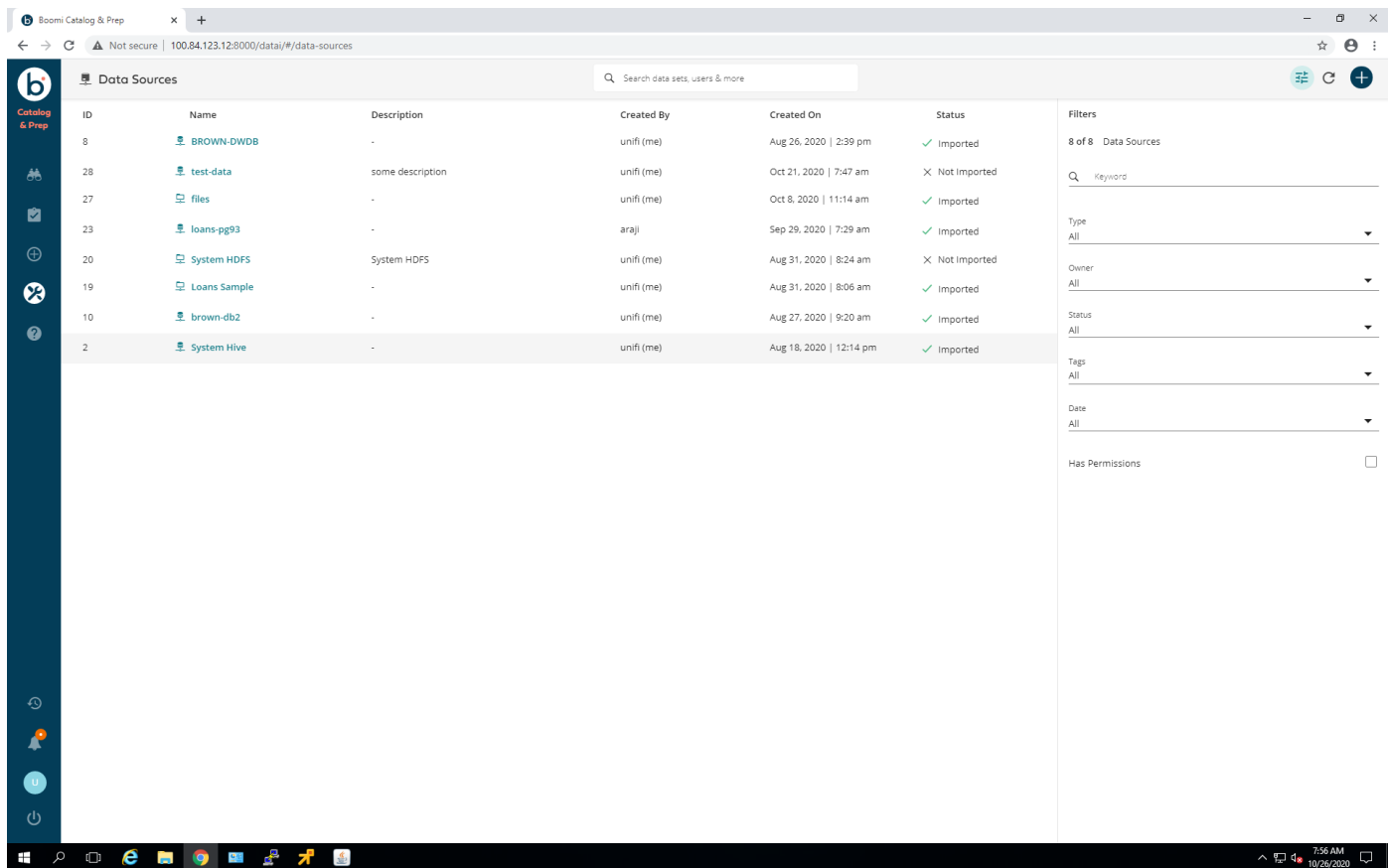


Figure 2. Data sources

Data sets

A data set is a subset of data from a data source. DCP users work with datasets to explore the available data, and to configure transform jobs.

Permissions can be granted to determine which DCP users can access each data set. The following user types can change data access permissions for a dataset:

- The data set owner
- The owner of the data source from which the data set was created
- A user with administrator privileges

Administrators typically control which adapters and data stores will be created and accessible in DCP, and which users can access each data store. Non-administrators can add a data source from a data store that was added by a DCP administrator if given permissions to that data store. If the non-administrator does not see a data store in the UI, then it has either not yet been added to DCP, or the user has not been granted permission.

Users and groups

For decades, the lack of clear definitions for sensitive data, and the limitations of data analytics platforms, have stifled interest in self-service analytics. Efforts that are leading to an overall reduction in many of the concerns that have been hindering self-service data services include:

- Defining what constitutes sensitive data like Personally Identifiable Information (PII)
- Industry guidelines for how to handle access

The Boomi DCP security model enables collaboration between users with different roles through a hierarchical, centralized permissions assignment structure. This model allows Boomi DCP administrators to create fine grain data access permissions. You can define which users, or classes of users, can see and use which columns of data.

To see and add users to the system:

1. Click the **Tools > Manage** menu.
2. Click **Permissions**.
3. Select **Access**.

The **Roles** screen appears. See [Roles](#).

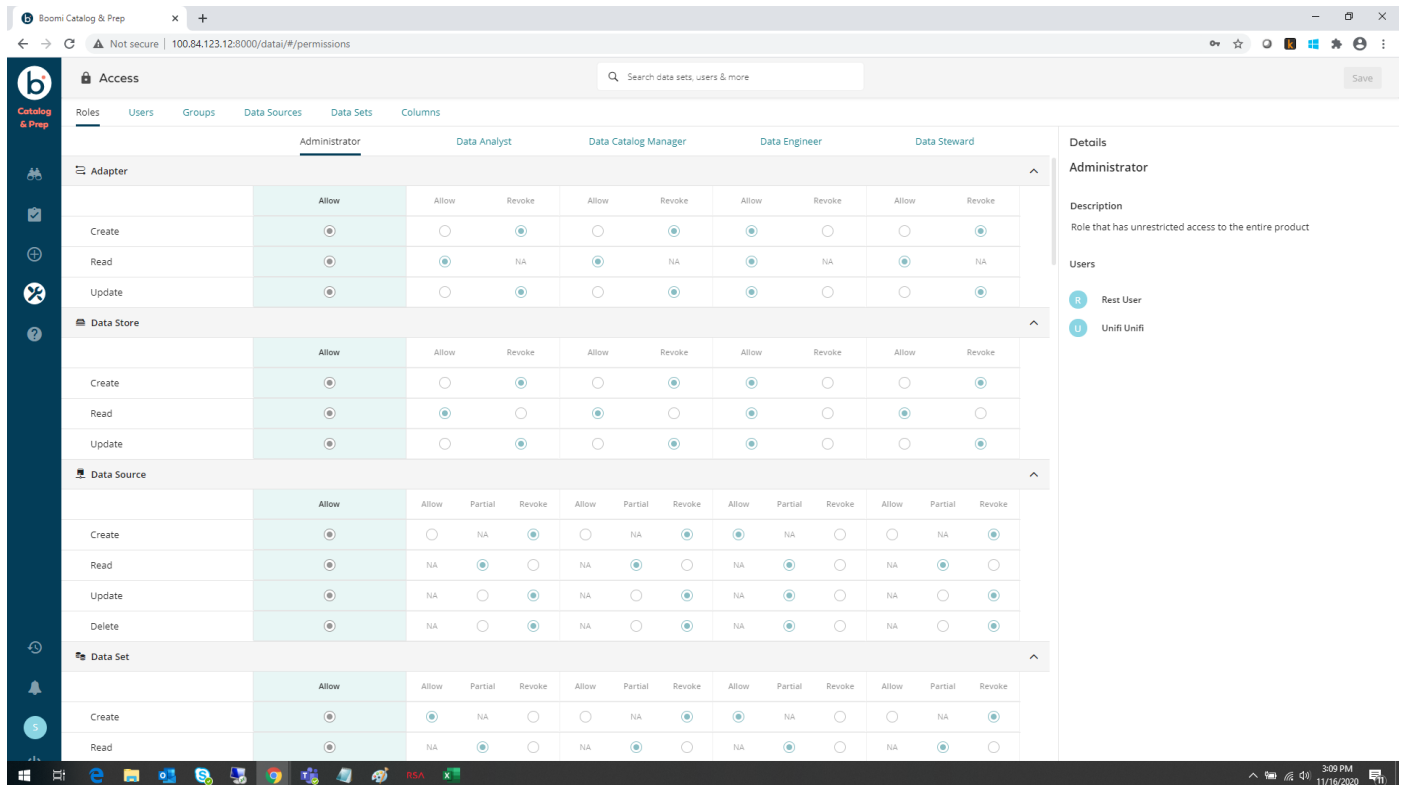


Figure 3. Roles

Roles

Five user roles are available to oversee data access and permissions management, as listed in [User roles](#).

Table 1. User roles

Role	Description
Administrator	This role has the greatest level of permissions on data, user accounts, and the behavior of the DCP application.
Data Analyst	This role acquires data from various sources and prepares it for data analysis, such as by creating transform jobs.
Data Steward	This role is responsible for data governance by administering data in compliance with policy or regulatory obligations. Users with this role see a customized view of the UI to fit their responsibilities.
Data Engineer	This role is responsible for developing architectures for building data pipelines. Users with this role see a customized view of the UI to fit their responsibilities.
Data Catalog Manager	This role is responsible for maintaining the inventory of all data assets in an organization. Users with this role see a customized view of the UI to fit their responsibilities.

These capabilities of these roles are influenced by the data hierarchy definition within DCP to establish what users can create and what users can access. The levels of data objects that get mapped to security roles are:

- Data Store

- Data Sources
- Data Sets

Only Administrators can enable and configure adapters.

See [Users](#) for a listing.

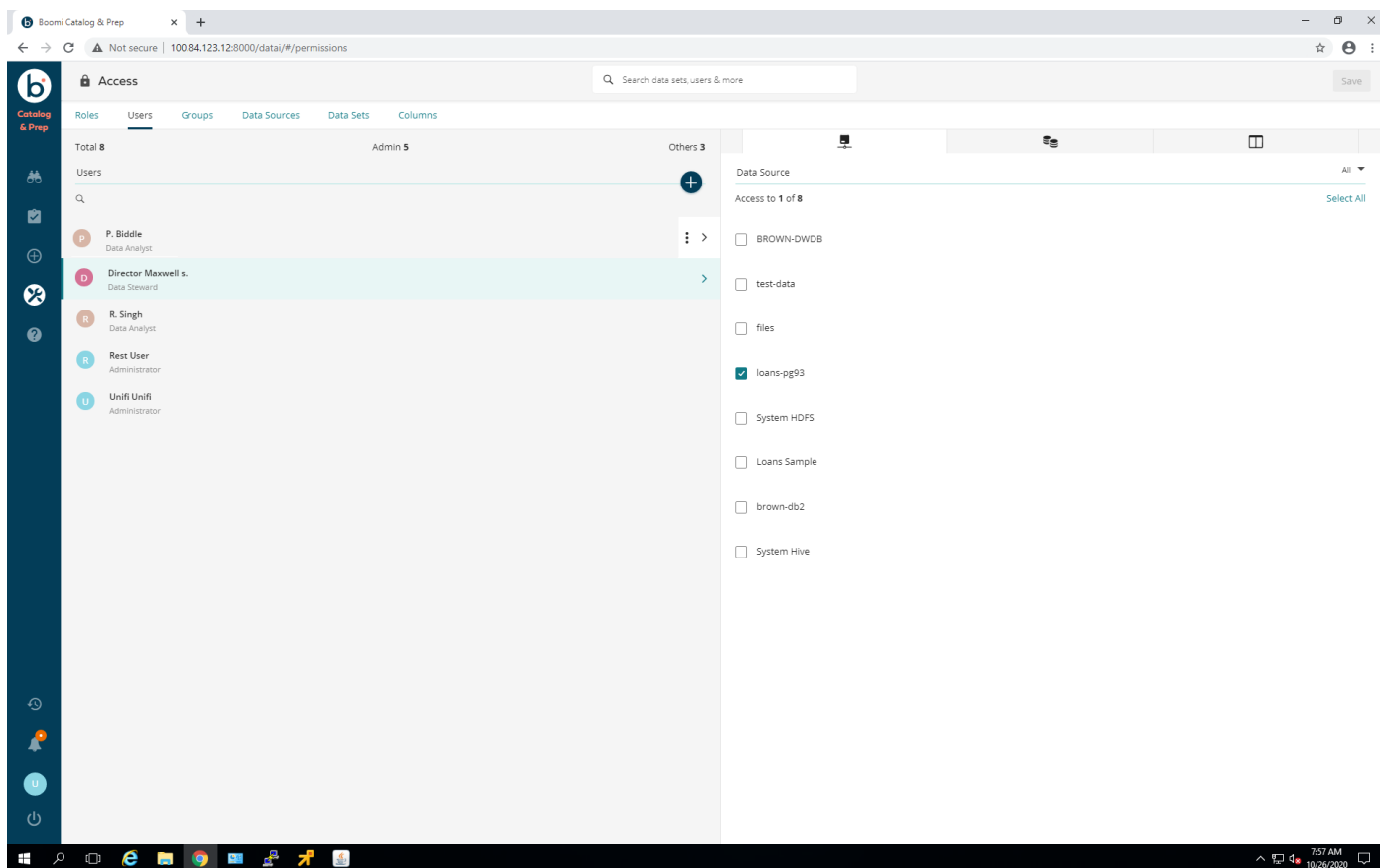


Figure 4. Users

Example users

The example includes five users with varying permissions, as listed in [Example users](#).

Table 2. Example users

User	Description
Unifi	The overall system administrator, with Administrator privileges
Rest User	A user created for running the RESTful API, with Administrator privileges
R. Singh	A data analyst, with privileges to view
Director Maxwell s.	A data steward
P. Biddle	A data analyst, with privileges to view

The **Access** page enables you to step through the various views describing:

- Roles
- Users
- Groups
- Data Sources
- Data Sets
- Columns

You can set permissions for all the objects; it can be as simple as selecting which users have access to which columns of data.

[Permissions](#) shows that R. Singh does not have privileges to see income.

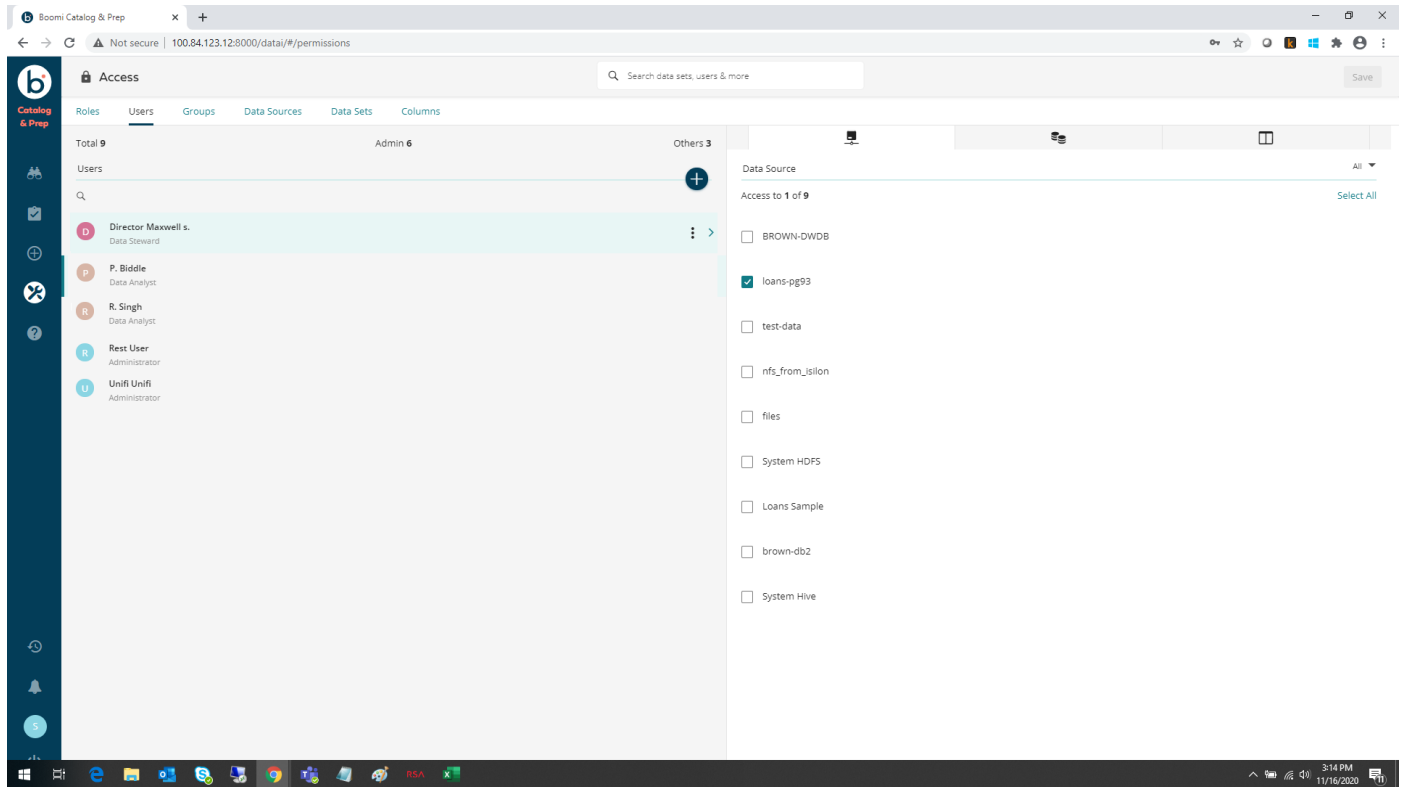


Figure 5. Permissions

Exploring data sets

About this task

Boomi DCP enables users to explore data without requiring them to directly access the data sources. To explore data sets:

Steps

1. Click the **Explore** icon (resembles a spyglass) on the left.
2. Select **Data Explorer**. See [Data set explorer](#).

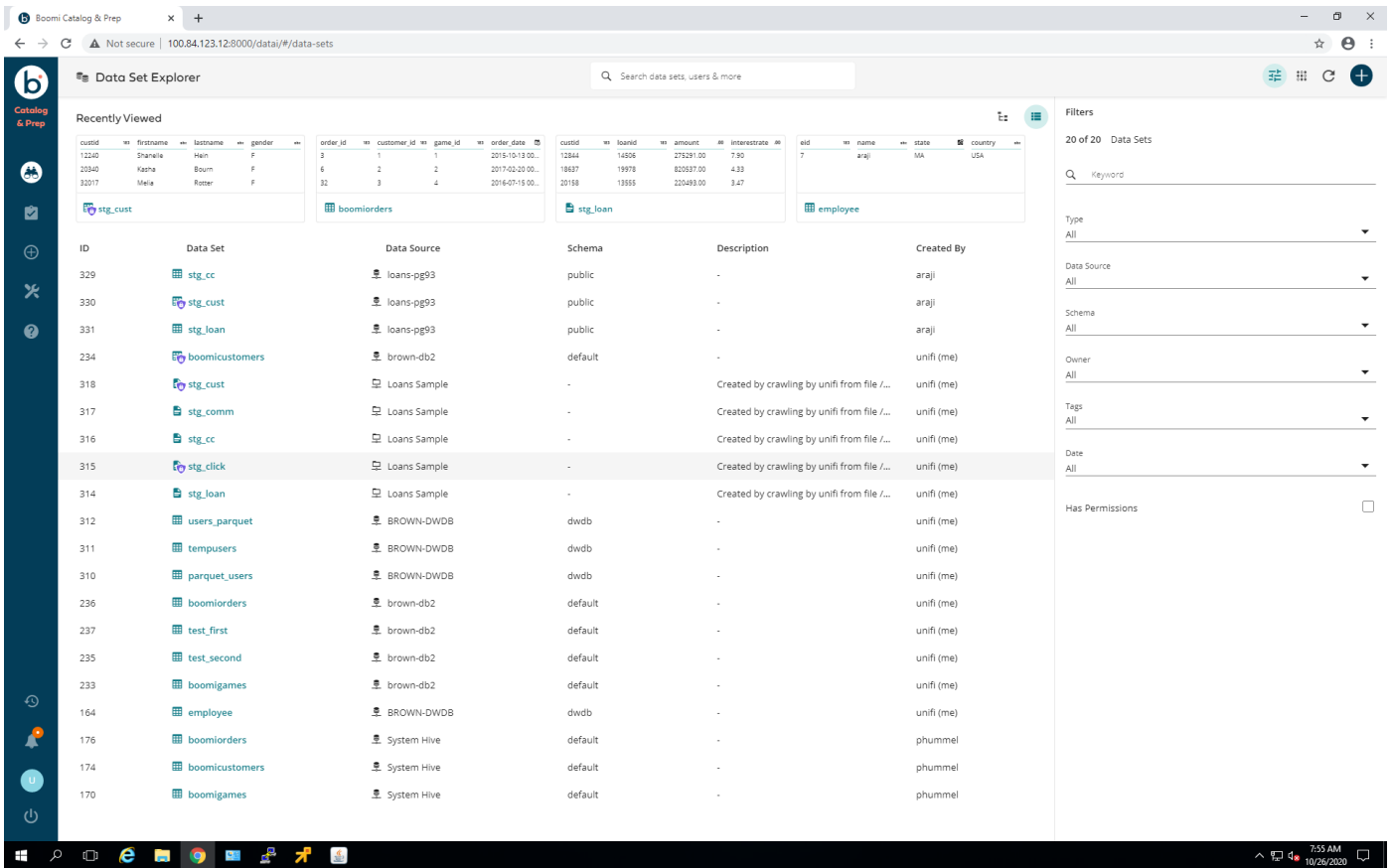


Figure 6. Data set explorer

3. Select any available data set to see example rows, and a graphical representation of the data.
4. Enter the following information:
 - Job Name
 - Job Description
 - Optional Tags
5. Click **Create**.

Data exploration is a powerful tool, and can reduce the time that it takes to create jobs.

Use case workloads

This document presents an example workflow in Boomi DCP. This example includes a fictitious loan department that has multiple data sources. Information about customers and loans that have been made resides in a PostgreSQL database. Information about customer communications resides in a file on a server. This workflow performs the following sequence:

1. Combines the information from these sources
2. Performs a data query
3. Filters the data
4. Transforms the data
5. Generates a result
6. Places the result into an HDFS data store

This example uses two data sources with multiple tables and a file to create an output. The output lists the number of customer communications for every successful loan.

Start the use case sequence

About this task

The user begins the use case sequence as follows:

Steps

1. Click the **Prepare** menu icon (a check mark).
2. Select **Jobs**, which displays a new page.
3. Click the **Add** icon (+) to start the jobs wizard.

The wizard contains seven stages for creating jobs. This example goes through each of the stages to show a sample job, and show the benefits for Boomi DCP users.

Select a data set

About this task

In this stage of the wizard, users can select any of the data sources that have been previously defined. The user selects a data set as follows:

Steps

1. Click **Select Data Set**. See [Select data set](#).

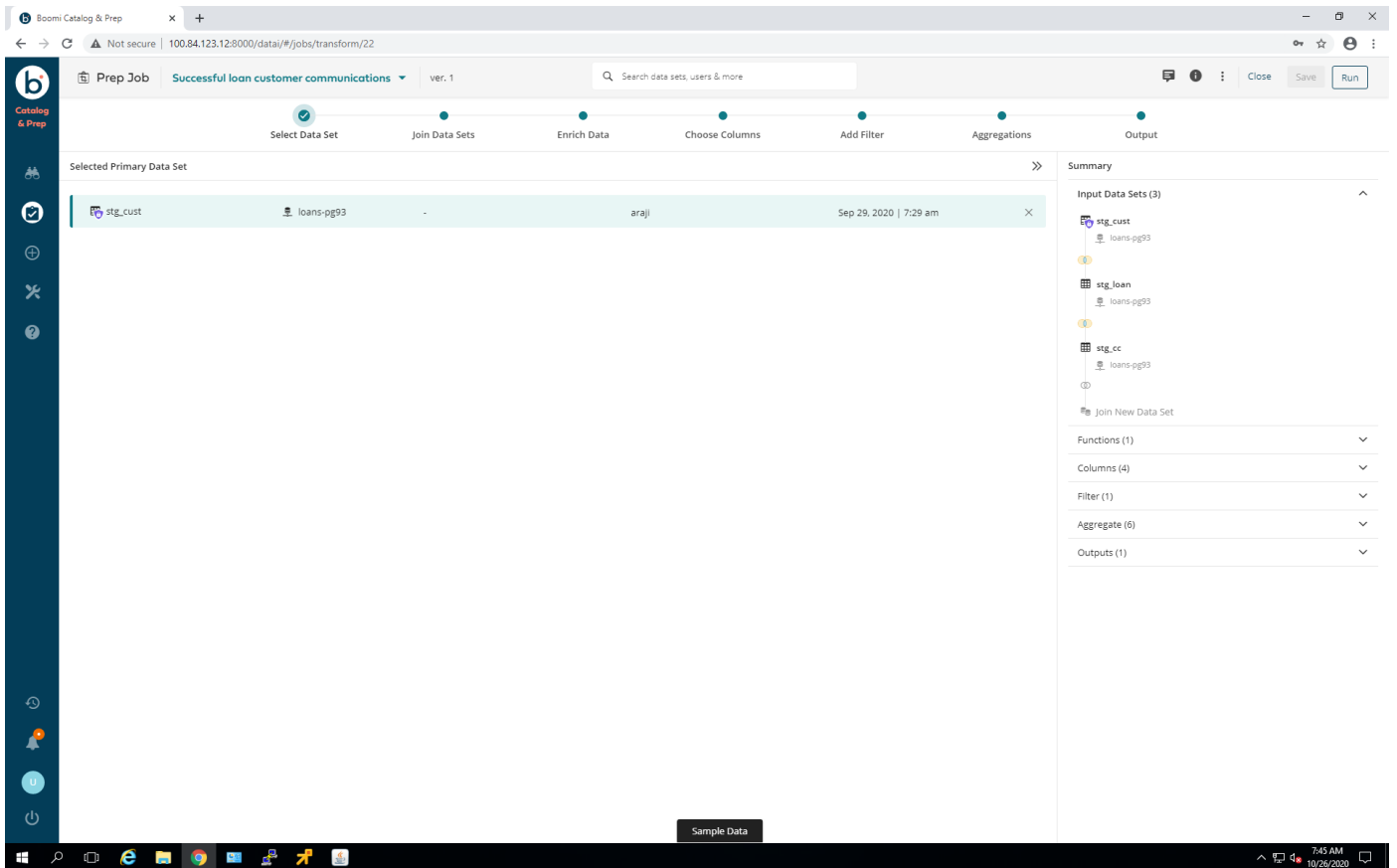


Figure 7. Select data set

2. Select **loans-pg9,3**, which is the PostgreSQL 9.3 database.
This database was previously created by adding a Data Store, and was not shown for brevity.
3. View the available data tables.
4. Select **stg_cust**, which is the customer table, as the primary data set.
5. At the bottom, click **Sample data** to display a preview of the data with which the user is working. See [Data preview](#).

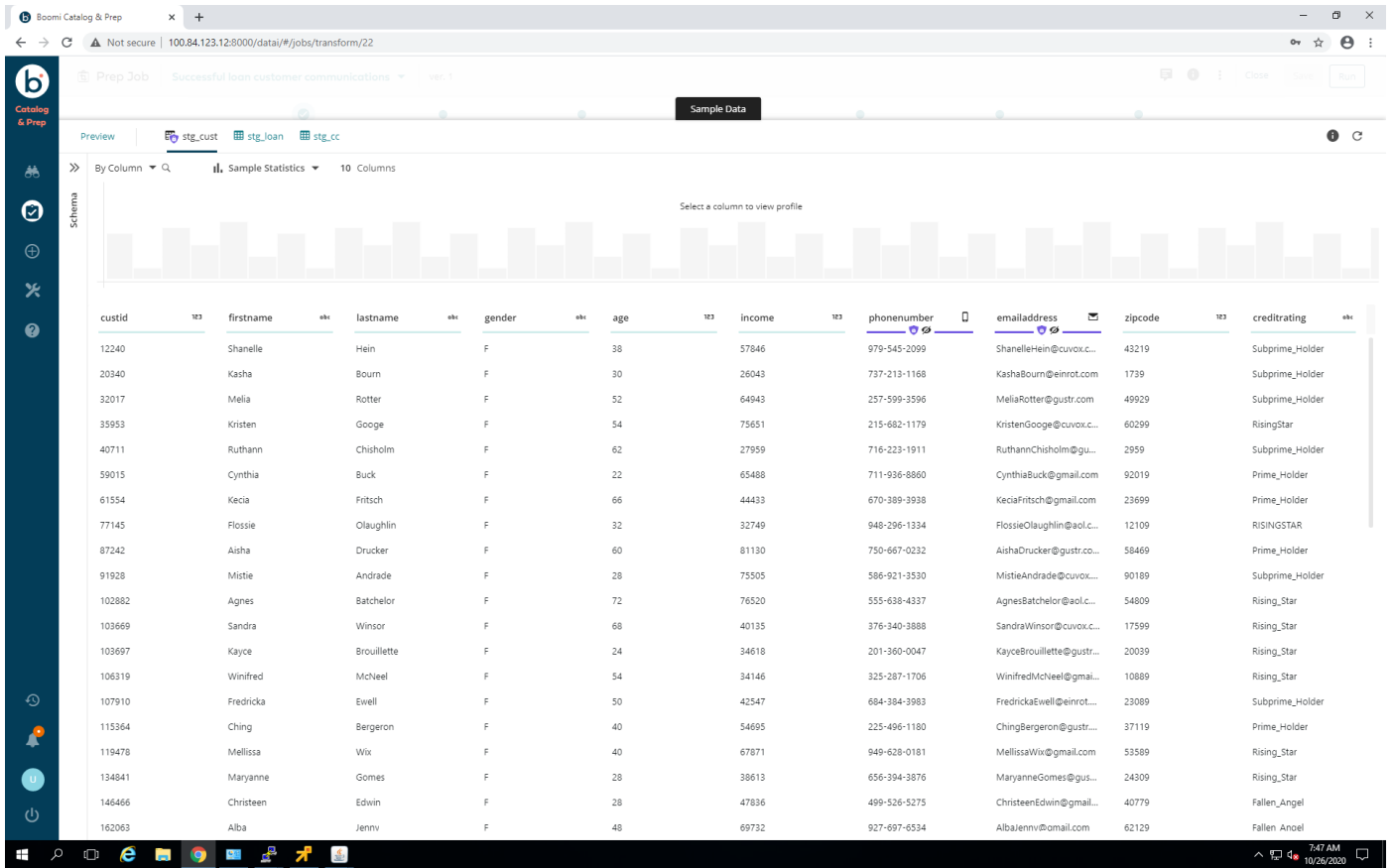


Figure 8. Data preview

At the top of the data preview, the user can see sample statistics about any column of selected data. These statistics are based on a subset of the data. The amount of data to be summarized can be configured. The viewing of sample rows of data and seeing a visualization of data are helpful, and are powerful tools.

Join data sets

About this task

To join data sets:

Steps

1. Click **Join Data Sets**. See [Join data sets](#).

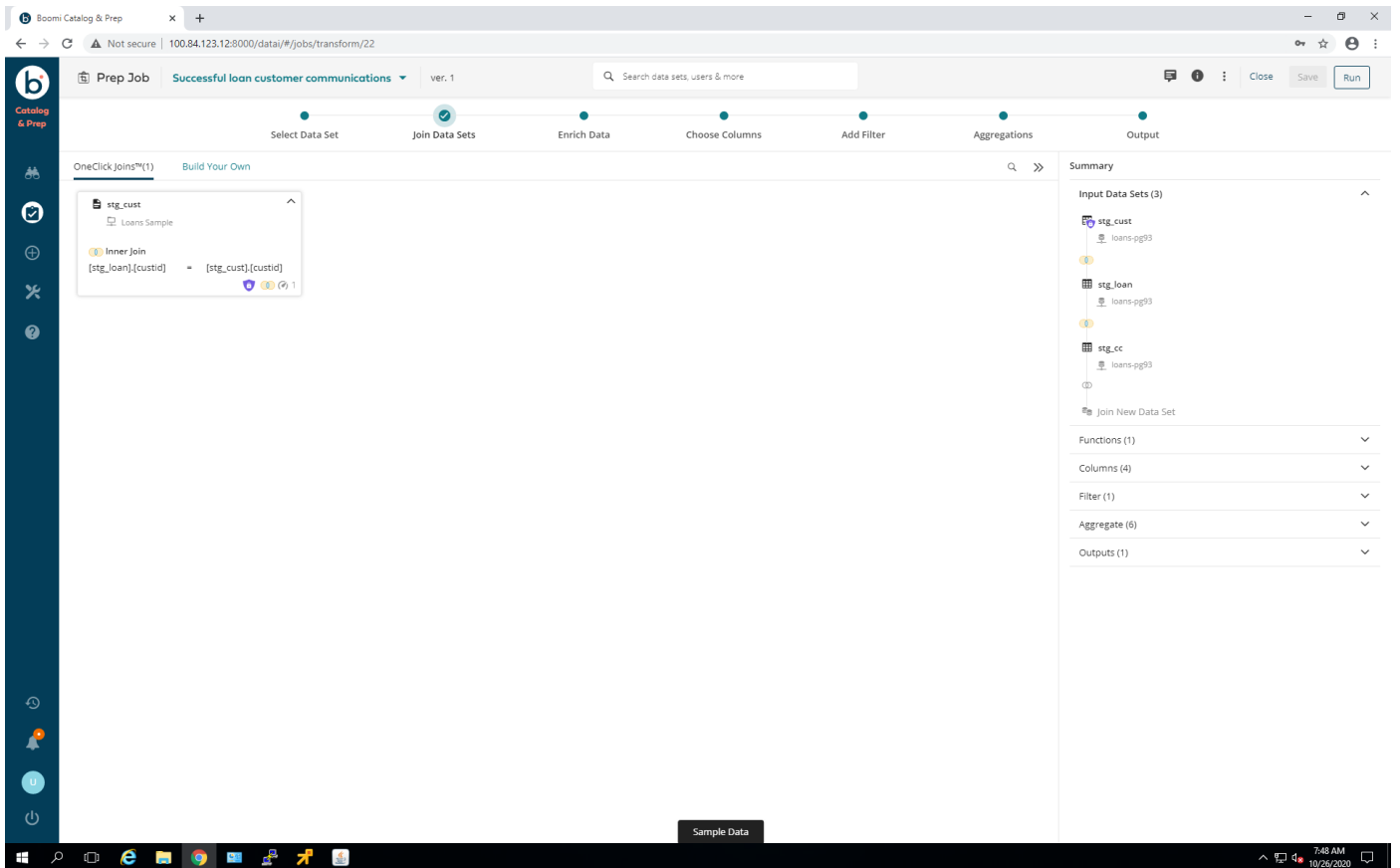


Figure 9. Join data sets

Boomi has multiple operators to allow for complex join operations. The joins themselves follow standard SQL functionality, such as which parts of the data to join. Boomi has additional functionality in this area to assist users. You can preview the data to be joined using the **Sample Data** pull-up menu.

Boomi also provides "OneClick Joins™", which leverages the underlying graph database that is loaded when DCP introspects data sets with foreign keys in databases or relationships between objects in Sales Force, for example.

2. Select additional data sets to be used in the query.
 3. Define how the data sets are joined, following standard SQL set definitions.
- Multiple such joins can be performed in this step.

Enrich the data

About this task

This step enables users to perform data transformations. This example merges the first and last names from the customer data table, and makes a single name. Boomi DCP comes with a rich set of predefined operators. Users can use these operators to create their own functions. This use case creates a data-enriching function as follows:

Steps

1. Click **Enrich Data**. See [Enrich data](#).

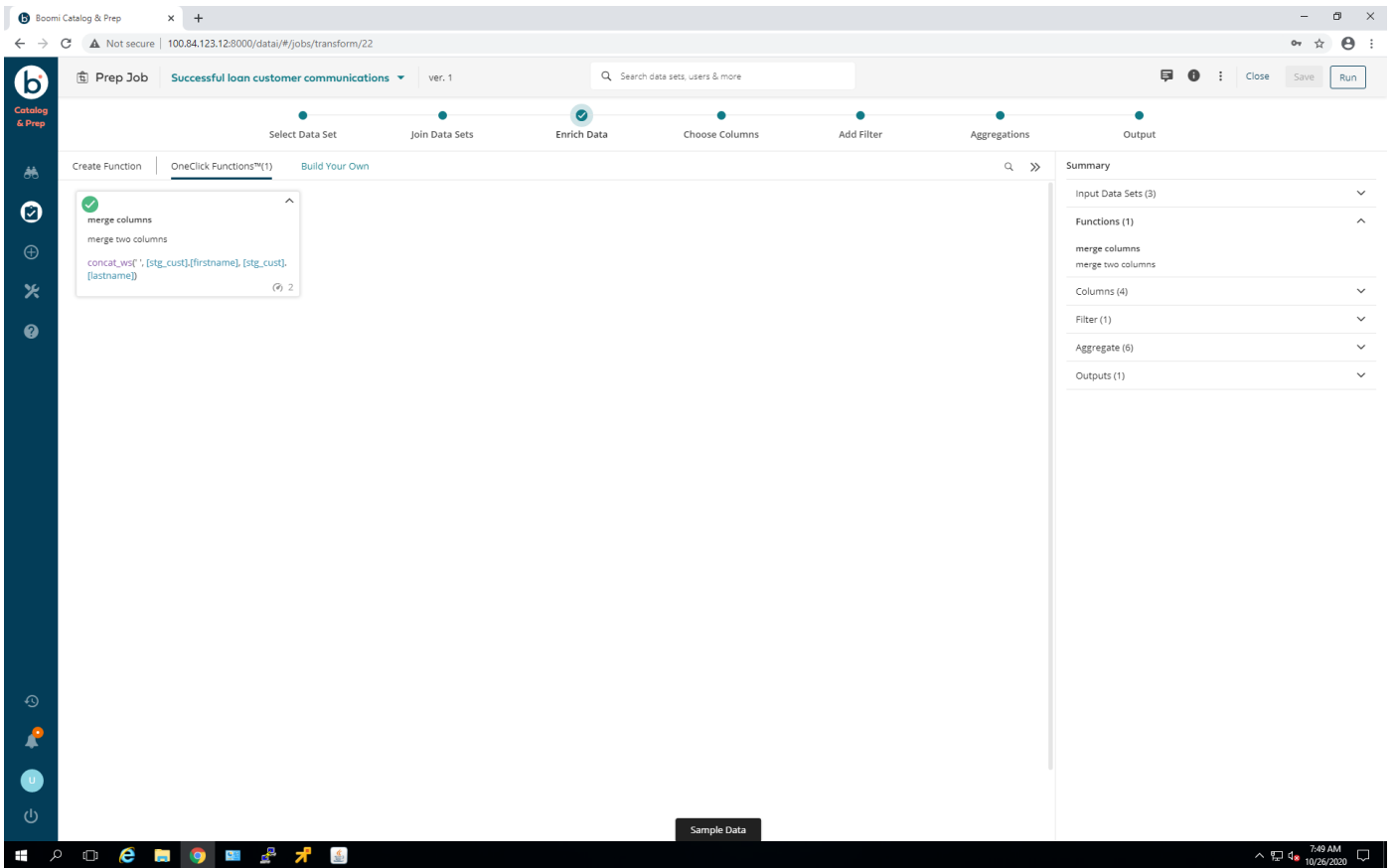


Figure 10. Enrich data

2. Click **Build Your Own** to define a function from scratch.
3. Provide a function name, **merge_columns**, and a description. See [Enrich data function](#).

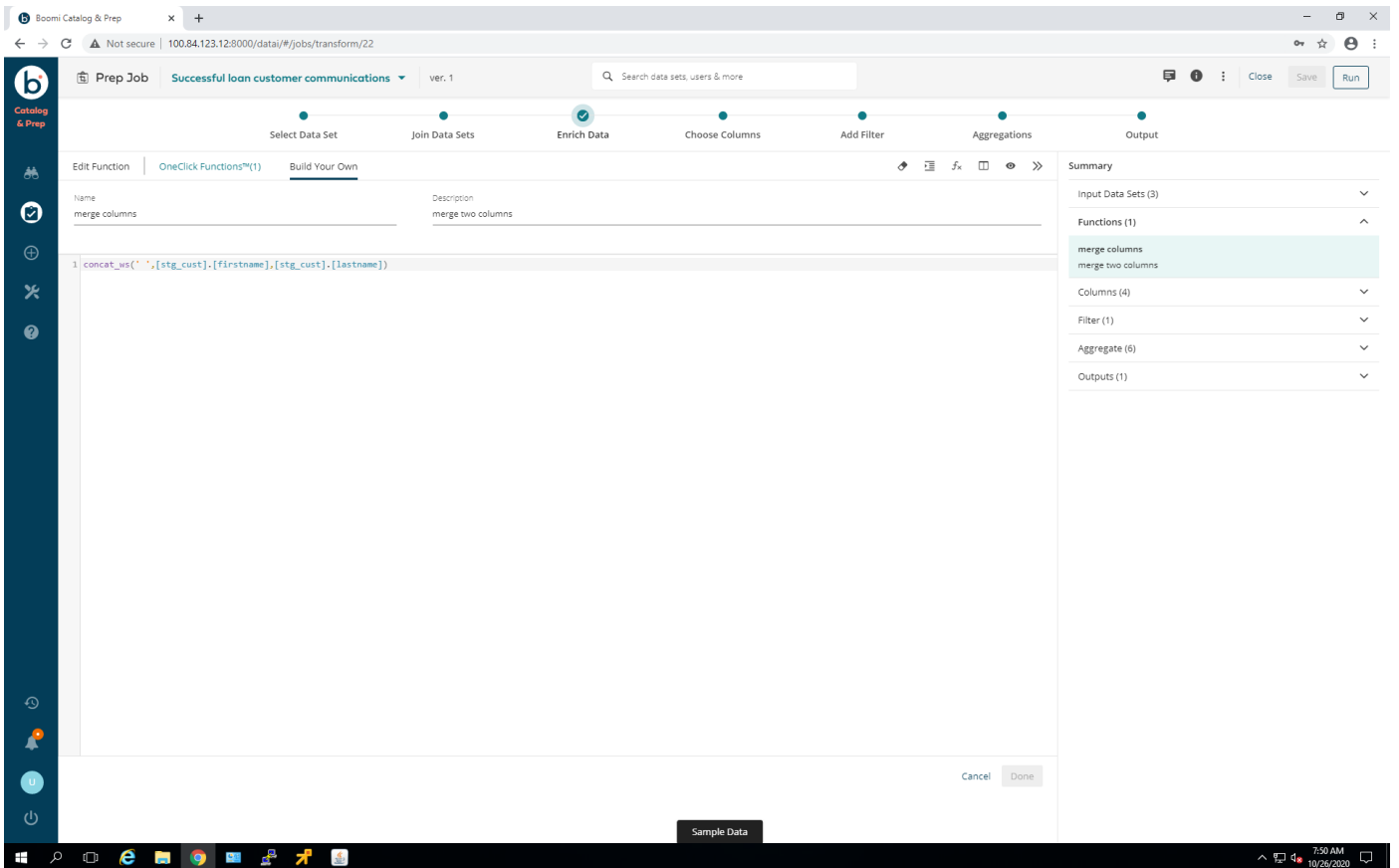


Figure 11. Enrich data function

4. Click the **Column** selection icon at the upper right.
Here you select one or more columns of data to use as inputs to your function.
5. At the top, click the **fx symbol** to display a list of integrated primitive functions.
These functions are standard for such data transformations.
6. On the right side, select **String > concat_ws**.
That selection concatenates elements from the two table columns, and combines them with a white space.
This wizard provides powerful tools for users. The documentation for all the functions is attached to each function in the UI. You can pull up sample data and see what your transformation will do, and much more.

Choose columns

About this task

This step enables users to select what columns, or pieces of data, will be part of the output. Choosing from the listed data columns and types creates the selection. For this use case:

Steps

1. Click **Choose Columns**. See [Choose columns](#).

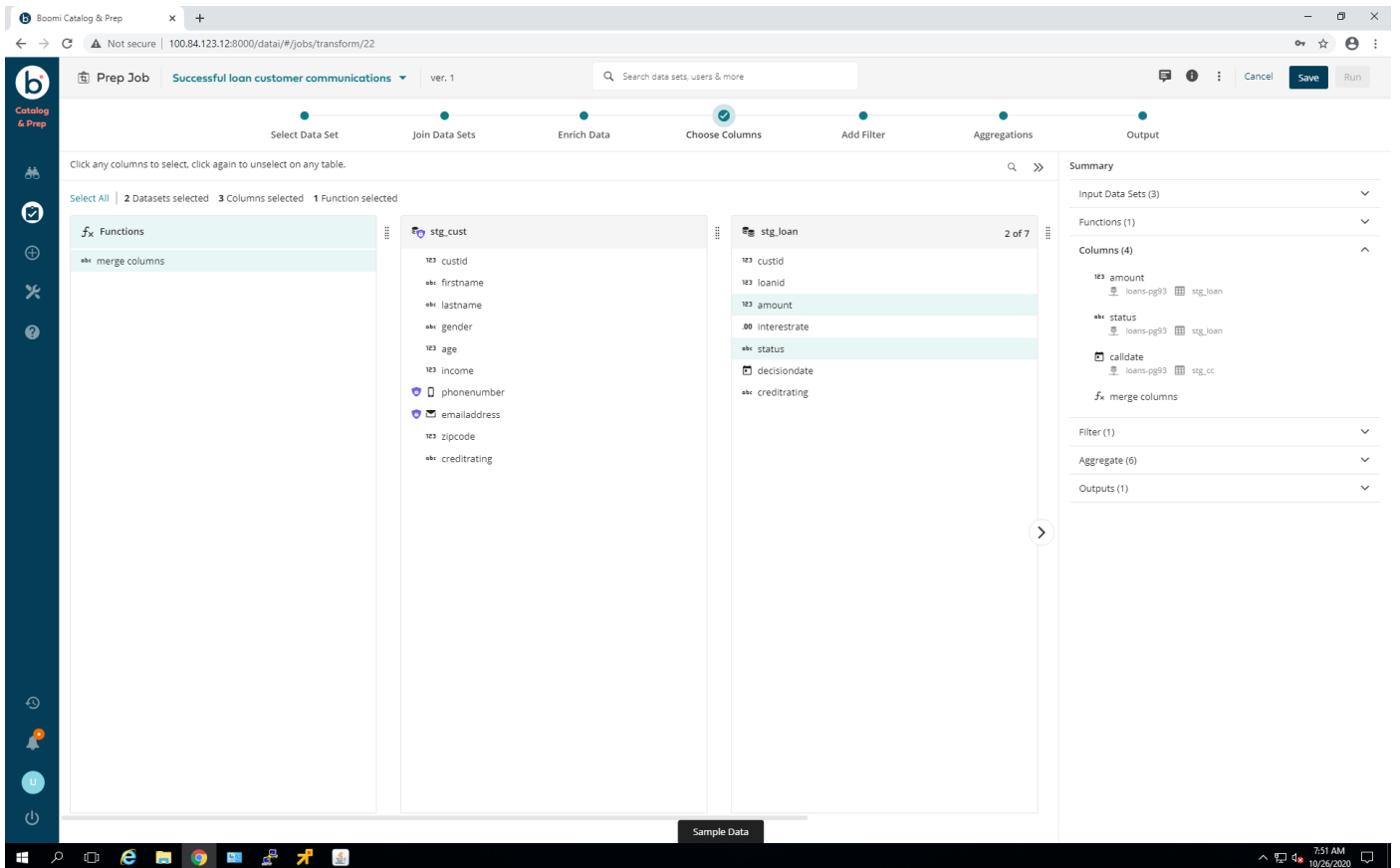


Figure 12. Choose columns

2. Select the output of the function created in [Enrich the data](#).
3. Select other columns of interest.

Add a filter

About this task

This step allows users to apply standard SQL filters to their job. Users can use all the standard grouping and selecting tools normally available. This use case is only interested in people with approved loans.

Steps

1. At the top, select **Add Filter**. See [Add a filter](#).

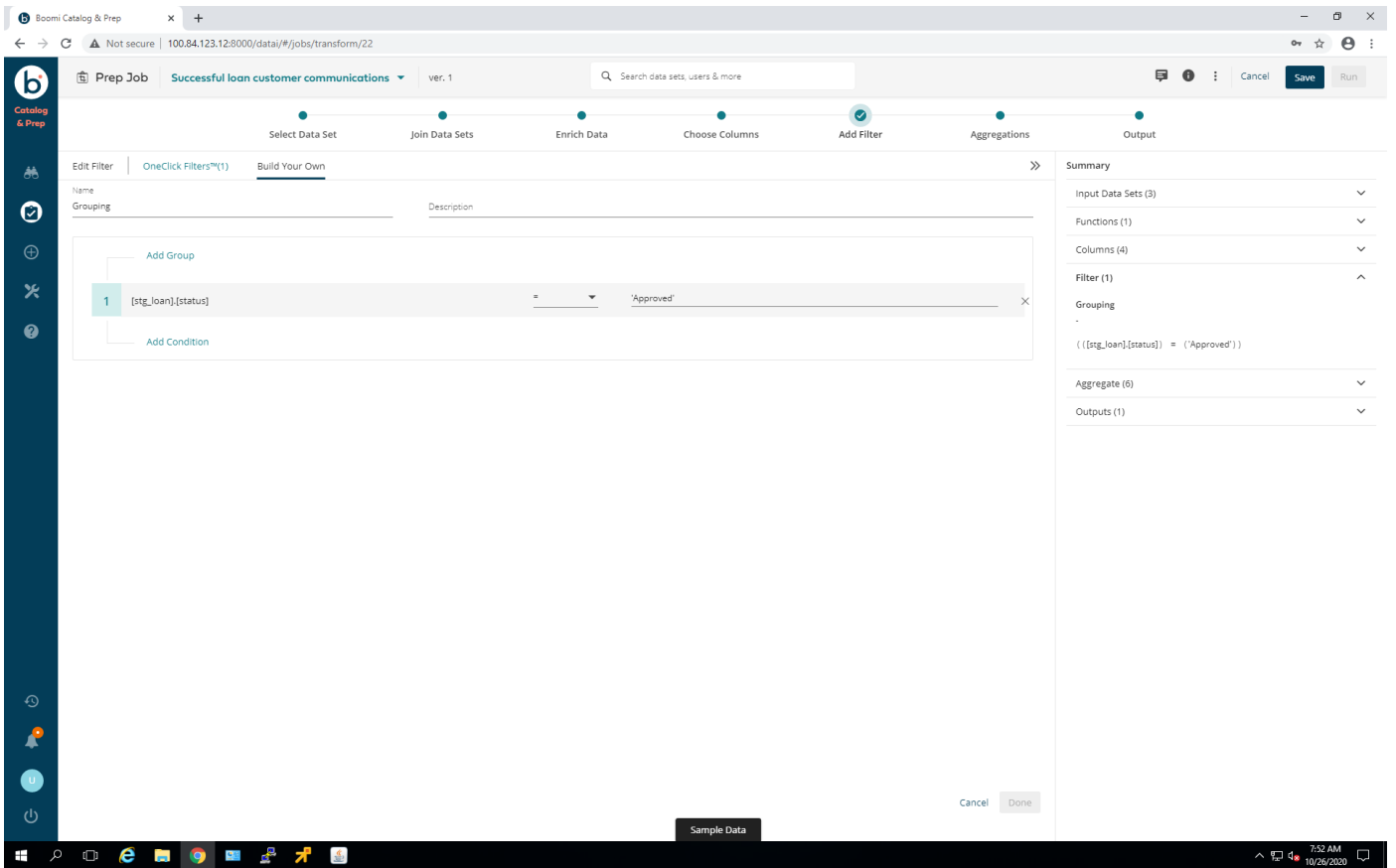


Figure 13. Add a filter

2. Click **Build Your Own**.
3. Click the **Add** icon (+), and then add the function text, **[stg_loan].[status] = 'Approved'**.

Boomi once again provides sample data and sample results that can be viewed by pulling up the **Sample Data** page from the bottom of the window.

Boomi has "OneClick Filters™", that also leverage the underlying graph database. When the data allows, it can be as simple as choosing one of the predefined filters created by Boomi using the graph data to get a valid and usable data output.

Aggregations

About this task

This step enables users to perform output functions on their selected data columns.

Steps

1. At the top, select **Aggregations**. See [Aggregations](#).

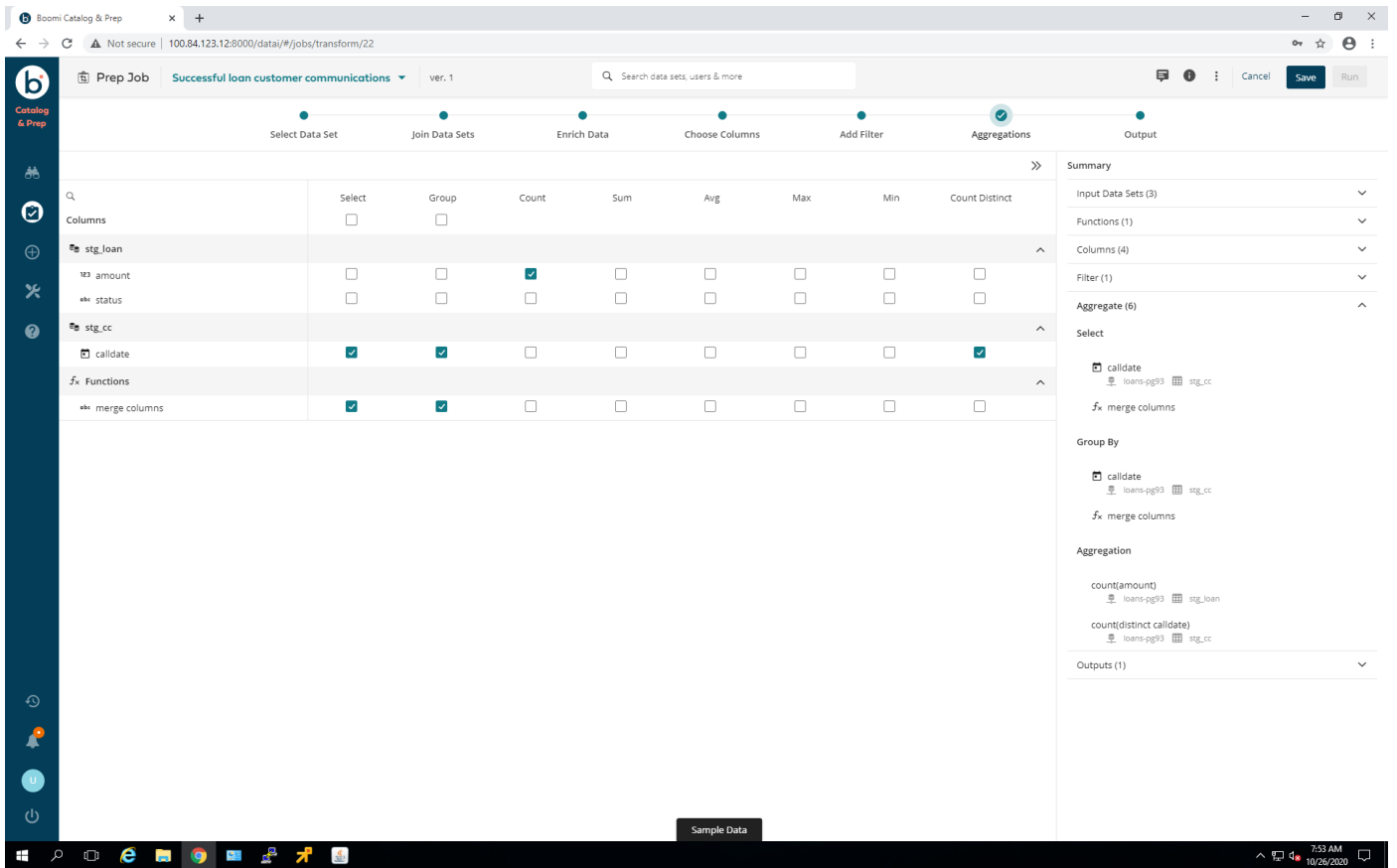


Figure 14. Aggregations

2. Click **Select** to select from that column or data source.
Each checked box results in the item being displayed in the output.
3. Click **Count** to provide a count of the selected column, and send it to the output.
Multiple aggregations are possible in this window. This use case example adds (counts) the amount of loans for each person.

Output

About this task

This step enables users to select any valid output data sinks that have been defined in Boomi DCP, or to create one or more data sinks to use. This use case example outputs the results to a file on an HDFS data store.

Steps

1. At the top, select **Output**. See [Output](#).

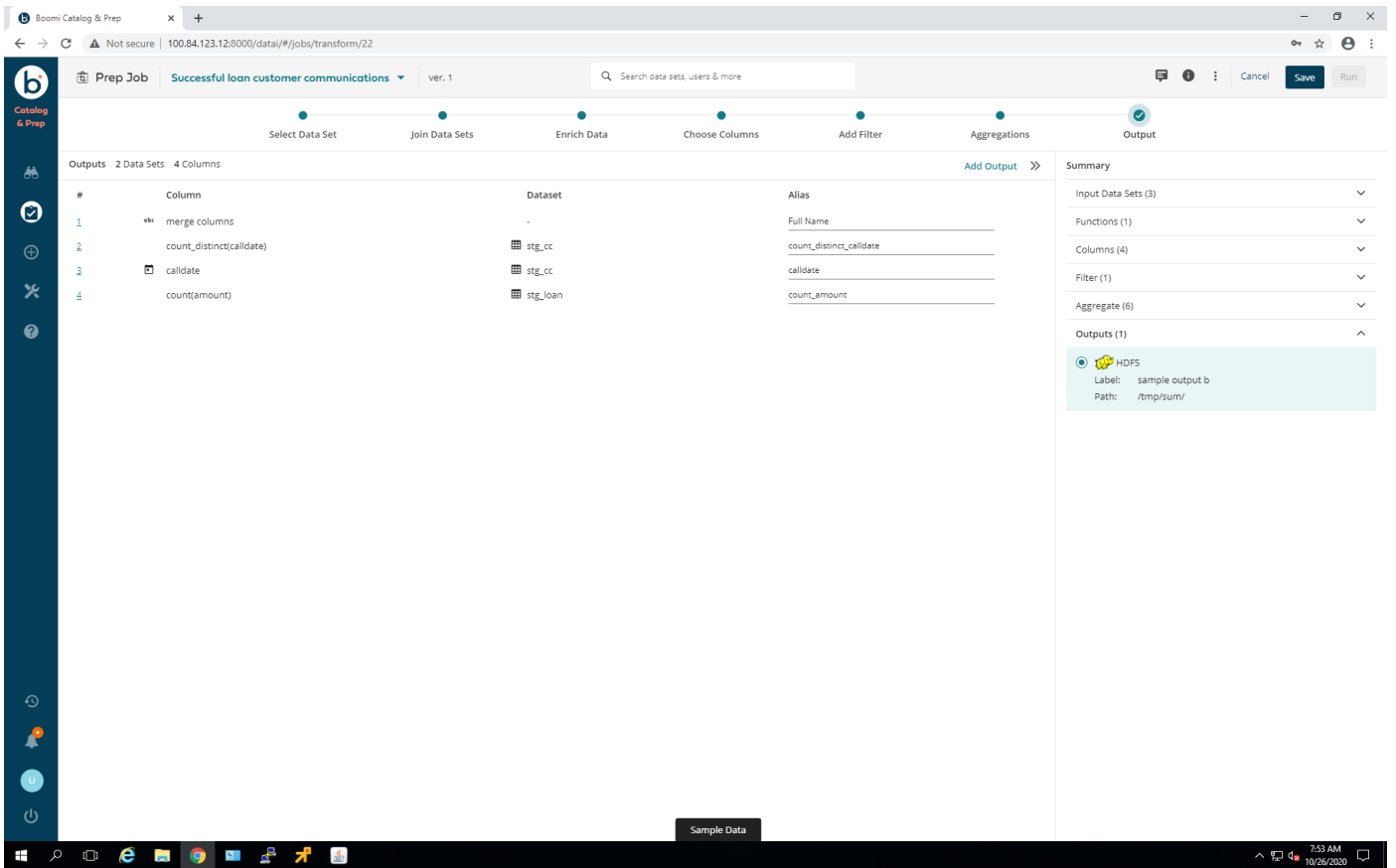


Figure 15. Output

- Note the following:
 - The human-readable name of each column
 - The original data source column
 - Where the data is written
- Select the output data sink. In this use case example, select the `tmp/sum` directory on the HDFS data store. The output enables users to select multiple types of output and locations.

Schedule a job

About this task

Boomi DCP includes an integrated scheduler to run jobs, which enables users to keep job results up to date on an ongoing basis. For example, a user may want to make loan summary data available every morning for loan officers. To schedule a job:

Steps

- Click the **Prepare** menu icon (a check mark).
- Click **Schedules**. See [Schedules](#).

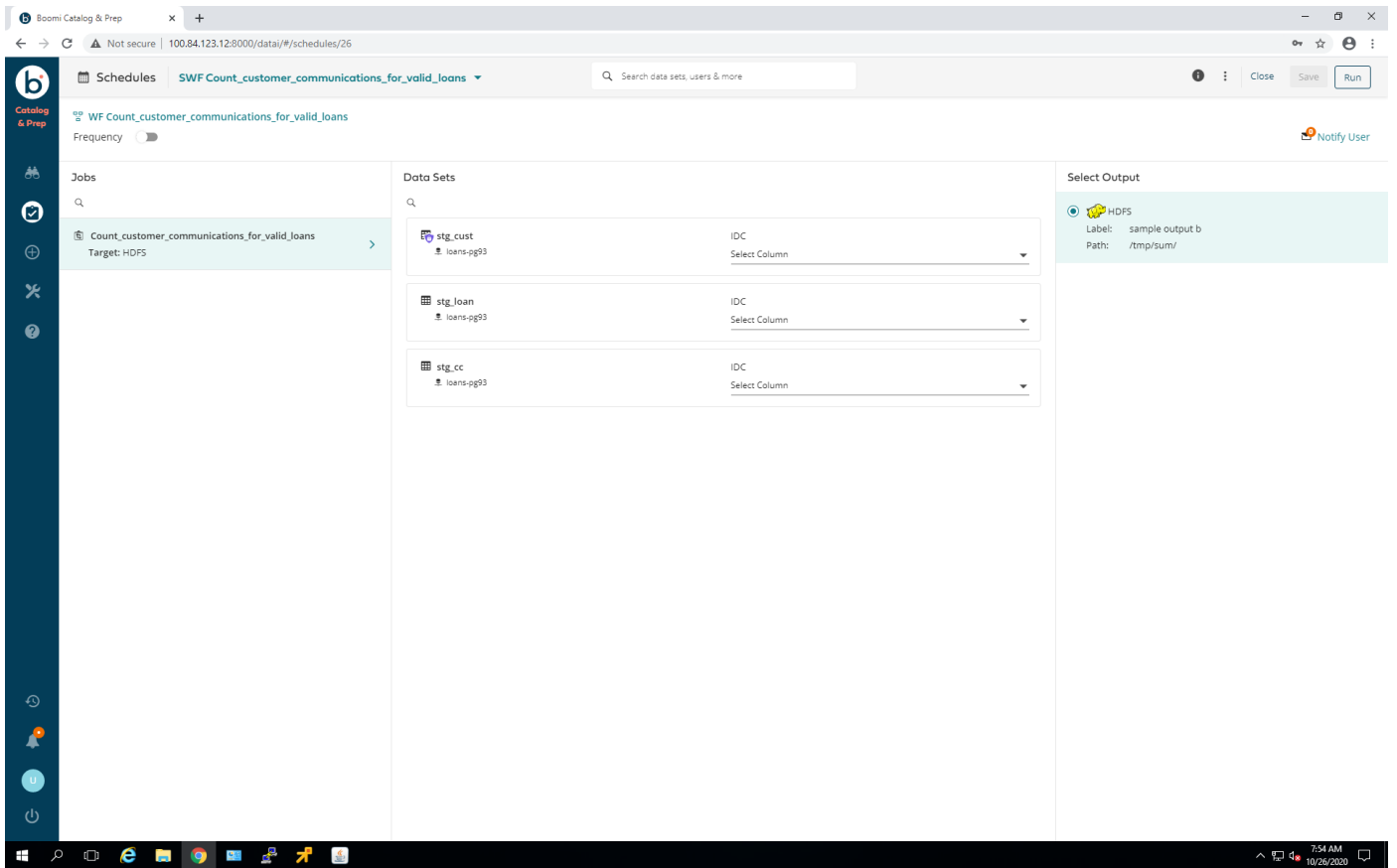


Figure 16. Schedules

3. Click the **Add** icon (+) to create a schedule for a job.
4. Fill out the form.
5. Click the **Create** button.

Once the schedule is created, users can add it to a job that they want to run it on. The scheduling is flexible, and easy to define.

RESTful API example

Boomi can be accessed using RESTful API requests to the server.

There are multiple ways to use the RESTful interface. Some examples include:

- Using an application similar to Postman.
- Using a programming language like Python.

Both have useful features.

The RESTful API can be used to do many of the things the Boomi web interface can, and some things it cannot. For example, there is no way to delete some objects, like datastores, from the web interface. This design choice was intentional by the Boomi team.

Using Postman to access Boomi

Using Postman, you can see what data is returned from individual calls to the RESTful API.

About this task

The graphical interface requires no coding to make requests.

Steps

1. Authenticate using an HTTP `POST` with Boomi and receive a token.
 - a. Use the URL, `http://<ipaddress>:8000/datai-api/get-jwt/`.
 - b. Set the body type to **Raw JSON**.
 - c. The body data should include the username and password fields in JSON format:

```
{
  "username": "user",
  "password": "pass"
}
```

2. Retrieve the token from the response JSON:

```
{
  "token":
  "eyJ0eXAiOiJKV1QiLCJhbGciOiJIUzI1NiJ9.eyJpYXQiOiJlMjMxMzUzZmJjQsImV4cCI6MTYwMzEzNTkyNCwiYWQiOiJlbnVzZXJuYWI1IjoidW5pZmkiLCJpcF9hZGRyZXNzIjoimTAuMTM1LjcyLjI2In0._SrmB0QAqcCXXkuxbSEqJ0yvCK7849AeI8XOFiA9QoU",
  "refresh_token":
  "da719a14cae742f7b1de24ff8a65cf6e8f58c8bdd73f4a25b34ec0299b8b5da6",
  "data": null,
  "last_login": null,
  "status": "SUCCESS"
}
```

3. Use the token in subsequent HTTP `GET`, `POST`, and `DELETE` operations.
 - a. Attempt to retrieve the users from `http://<ipaddress>:8000/datai-api/users/`.
 - b. In a `GET` request, select the **Authorization** tab.
 - c. Choose type **Bearer Token**, and then place the value of the token in the **Edit** field.
 - d. **Send** the request.
 - e. Examine the output.

The token has a short life span. It is only valid for approximately 5 minutes. If you use an expired token, you get an error back from the REST call:

```
{
  "detail": "Token signature has expired",
  "status_code": 401
}
```

This limitation makes using the Postman interface cumbersome.

The Postman utility can be found at: <https://www.postman.com/>.

Using a programming language to access Boomi

Another way to access the API is by writing code in Python or a similar scripting language. This method makes it easier to log in and get a new token when the script is run.

About this task

If the script is more interactive and long-running, the expired token can be caught and a new token requested. Then the operation can be performed.

Steps

1. Create a script to perform the operation. For example:

```
#!/usr/bin/python3

import requests
import json
```

```

bodyStr = json.dumps({ "username" : "user", "password" : "pass"})

# get the required token for making requests to the REST API
response = requests.post("http://127.0.0.1:8000/datai-api/get-jwt/", data=bodyStr,
headers={"Content-Type": "application/json"})

if response.status_code == 200:
    # success
    responseDict = json.loads(response.text)

    # setup the requests header to pass the token
    headers = {"Content-Type": "application/json", "Authorization": "Bearer " +
responseDict['token']}
    usersResponse = requests.get("http://100.84.123.12:8000/datai-api/users/",
headers=headers)

    usersJson = json.loads(usersResponse.text)

    for user in usersJson:
        print(user['username'])
else:
    print("Error: " + response.text)

```

2. Run the script.
3. Examine the output.

Integrating Boomi DCP with Hadoop

Boomi DCP can be used in stand-alone mode, or with Hadoop. Adding Boomi DCP to an existing Hadoop cluster adds significant functionality.

Topics:

- [Overview](#)
- [Boomi server configuration](#)
- [Compatible Dell EMC hardware](#)

Overview

Boomi provides users with a simplified ETL (extract, transform, load) experience. The all-in-one tool has significant value to organizations trying to leverage their Hadoop investments.

Boomi server configuration

There are no hard specifications for the type or size of server that is used for Boomi DCP. The hardware limits how quickly data can be ingested, cataloged, and used. The Boomi DCP server should conform to the general specifications listed in [General specifications](#). Dell Technologies recommends that you buy a server matching the Infrastructure Node configurations that are defined in the **Hardware Infrastructure** section of the [Dell Technologies Cloudera CDP Data Center on Dell EMC Infrastructure Reference Architecture](#).

The Boomi server requires an attached Cloudera cluster to perform ETL and transformation operations. The catalog portion of Boomi only requires its backend PostgreSQL database. The catalog functionality still provides great user value, and can reduce user time to value greatly.

General specifications

General specifications for the Boomi DCP server are listed in [Boomi DCP server general specifications](#).

Table 3. Boomi DCP server general specifications

Component	Minimum	Preference
Operating system	An enterprise Linux distribution	RHEL 7.x, CentOS 7.x, or Amazon Linux 2.0.
RAM	At least 64 GB	128 GB or more (memory to vCPU ratio of 8 GB per vCPU)
Processor	x86 processor with at least eight cores	x86 processor with 16 or more cores
Storage	500 GB	More than 1 TB, for the PostgreSQL database and other parts of the application stack that are installed

Disk specifications

Disk specifications for the Boomi DCP server are listed in [Boomi DCP server disk partitions](#).

Table 4. Boomi DCP server disk partitions

Mount point	Size
/usr/local	500 GB or more
/tmp/	50 GB or more
/home/unifi	50 GB or more
/opt/	50 GB or more
/var/log/	50 GB or more

Software specifications

Software specifications for the Boomi DCP server are listed in [Boomi DCP server software specifications](#). All software is installed during the DCP installation process.

Table 5. Boomi DCP server software specifications

Package	Install location
PostgreSQL	/usr/local/pgsql
Python	/usr/local/lib/python3.8
Python setuptools	/usr/local/unifi_virtualenv/lib/python3.8/site-packages/
pip	/usr/local/bin/pip3.8
virtualenv	/usr/local/bin/virtualenv
Redis	/usr/local/bin/redis-server
Solr	/opt/solr
Java	/opt/jdk1.8.0_131
Nginx	/usr/local/nginx
Spark binaries	/usr/local/spark
gcc, development tools (devel), yum	/usr/bin/

Hadoop specifications

Boomi DCP is compatible with both Hortonworks and Cloudera Hadoop as data sources and data sinks, but only Cloudera Hadoop for Boomi Data Prep backend (ETL operations). The Boomi node can be added to Cloudera at the time of cluster install, or later. Boomi has a multitude of adapters that enable read and write operations. The Hadoop services that are used in this manner include those services that are listed in [Hadoop services](#).

Table 6. Hadoop services

Service	Versions
Cloudera Distribution for Apache Hadoop (CDH) HDFS	5.3.0, 5.7, 5.12, 6.0
CDH Hive	5.3.0, 5.7, 5.12, 6.0
Hortonworks Data Platform (HDP) HDFS	2.4, 2.6
HDP Hive	2.4, 2.6
MapR HDFS	4.0.2
MapR Hive	4.0.2

Table 6. Hadoop services (continued)

Service	Versions
Apache HDFS	2.5.0, 2.6.0, 2.7.1, 2.7.3
Apache Hive	0.13.1, 1.1.0, 1.2.1
Amazon EMR HDFS	5.8.0, 5.11
Amazon EMR Hive	5.8.0, 5.11

Required Cloudera Hadoop services on the Boomi DCP node that are used for Boomi Data Prep backend (ETL operations) include:

- HDFS Gateway
- Hive Gateway
- Sqoop 1 Client Gateway
- YARN (MR2 included) Gateway

Required services on the cluster data nodes include:

- HDFS DataNode
- Hive Gateway
- YARN (MR2 included) NodeManager

All these services can be defined either at the time the cluster is installed, or later. Conceptually, the Boomi node becomes a member of the cluster for access, but not as a Hadoop storage or processing node. Boomi uses the gateway functions to perform the ETL operations.

Boomi creates jobs on the Hadoop cluster to:

1. Read the data into HDFS.
2. Place it in Hive.
3. Perform various operations.
4. Write the result out to the selected destination.

Data processing does not occur on the Boomi node, but on the source, destination, and Hadoop cluster.

Compatible Dell EMC hardware

Boomi DCP is compatible with a wide range of hardware, both to run on and to access as a data store. Dell Technologies used a Dell EMC PowerEdge R640 server with 192 GB of memory to test the software.

Boomi DCP can likely run on one of the same types of Dell EMC server nodes that you already have in your cluster today. This compatibility simplifies ordering, maintenance, and planning operations in your environment. The [Dell Technologies Cloudera CDP Data Center on Dell EMC Hardware Reference Architecture](#) includes sample Infrastructure Nodes suitable for Boomi.

Solution architecture

Boomi DCP helps enterprises quickly and accurately discover, acquire, and transform data so that it can be combined with existing data. It also enables data to be analyzed using business intelligence and visualization tools to extract business insights.

Topics:

- [Overview](#)
- [Hadoop cluster](#)
- [Boomi DCP server node functions](#)
- [Networking](#)
- [Metadata catalog](#)

Overview

Boomi DCP is a Dell solution for self-service data and business intelligence that offers the following functionality:

- Discover, organize, and catalog your data, wherever it is located, in whatever form it is in.
- AI tools that simplify the jobs to prepare your data with a few clicks.
- Secure, protect, and manage access to your data by role, and by dataset, row, and column.

With Boomi DCP, business analysts can select and integrate their data sets without having to write code or involve IT personnel. This functionality enables analysts to pursue "what if" scenarios with the data, and develop business insights quicker than with traditional, hand-coded programming.

Boomi DCP enables business and data analysts, data engineers, data scientists, and other lines of business (LOB) data consumers to discover and understand relevant data sets in a self-service mode. Users can extract business value under required levels of security and governance.

Hadoop cluster

Boomi DCP runs on a separate server from the Hadoop nodes. It should reside on the same network segment as the Hadoop clusters, to minimize latency. The server requires both edge and cluster network access in order to accommodate requests and access the data.

The Boomi server can be added to an existing Hadoop cluster or in a new installation. Adding it to a cluster can be done using the Hadoop user interface, and is as easy as adding any other node. A few connectors are added to the Boomi interface.

Boomi can discover the data sources in the cluster through the Unifi portion of the stack. It can connect to the existing Hadoop cluster using one of the provided plug-ins, such as the HDFS plug-in. Data discovery and ingestion can occur once the connection is established. As the data source is added, Boomi can walk through the data and generate catalog information for later use.

The robust security model enables fine-grained control over which portions of data users can access.

Boomi DCP server node functions

The Boomi DCP node provides for data discovery from various sources:

- SQL and NoSQL databases
- Hadoop
- Amazon S3
- Locally attached storage such as NFS

- Data streams such as Twitter, Facebook, and more.

Boomi DCP creates a metadata data store from cataloged data. Once data stores have been added, users can:

- Browse data samples.
- Generate statistical views of the data.
- Create jobs using the Boomi interface to perform ETL without writing code.

Boomi also provides a scheduling service to perform such jobs on an ongoing basis.

Networking

Boomi DCP requires a network path to the data and cluster. This path depends on how the Hadoop cluster network is configured at the customer site.

For security, the Boomi DCP node should run a firewall to limit access to services. Rules for ports 8000 and 8443 must be added to enable access to the web interface and RESTful API.

Metadata catalog

A metadata catalog is a set of relational tables. They contain information about data and how to convert data from nonrelational to relational formats. The Boomi server accesses the information that is stored in these catalogs to perform later operations.

Boomi maintains data sets that are the metadata representation of the data source. As data sources are added to Boomi using a connector to an external database, file or application discovery is performed. During discovery, Boomi categorizes and examines the data in place, and then copies a sample of the data to itself. Data remains on the source. Only sample data are loaded temporarily to perform analysis on the data, and metadata are stored for Boomi functionality.

The metadata catalog has two components:

- Technical metadata, which is gleaned from the source during discovery
- Business metadata, which users add collaboratively based on what the data means to the business

The technical data consists of:

- Table names
- Column names
- Data types
- Key constraints
- DBA comments to the columns

If the source is a file system (XML, JSON, Parquet files, and so on), Boomi creates a schema for the file. It then "infers" the data type information for the metadata.

Boomi can also identify sensitive or PII information from the source, which is classified and tagged automatically. The internal security settings hide content unless an individual has the required privileges to see it.

All this information is searchable using the Boomi web interface.

Annotation

The second part of the metadata catalog is the business metadata. It is built either manually or from crowdsourcing and collaboration with others who have access to the data. Business metadata helps users understand how information from different sources can be paired together for data analytics tasks.

Boomi has a collaborative environment, where users with access to the data can add understanding and rate the importance of the data. This environment includes the data dictionary and business glossaries.

Curation

Curation is the process of managing the data and its usefulness. The Boomi interface makes it easier to perform curation. It uses a graph database and integrated AI to analyze the data. Boomi enables users with little technical experience to manipulate and use the data in many ways. It helps draw connections between data sets and create relationships for business purposes.

Boomi also enables jobs to be set up and scheduled to import new data from the configured sources. Jobs can also work with the expanded data sets and perform transforms and additional annotation, to add more business value. These jobs can be joined to form a multiple-step process for dealing with the data.

Tagging

Tags can be used in Boomi to group different objects together under a common theme. Some of the objects in Boomi that can be tagged include:

- Data sets
- Data sources
- Jobs
- Schedules
- Users
- Workflows

Tags can consist of lowercase alphanumeric characters, dashes (-), and underscores (_). A space ends a tag when entering a tag through the web interface. Continuing to type in the entry starts a new tag.

The Boomi web and RESTful interfaces each have search features that allow the use of tags to find all related objects.

Conclusions

Significant investments have been made over the last decade to organize the data assets of organizations into data lakes that are easier to catalog and consume. Digital Transformation results in a host of new business applications. Many of them create data silos that have not been brought into a central data store or data lake. As a result, the complexity of the data landscape negatively impacts the productivity of business analysts. Finding and understanding data remains a time-consuming and experimental activity.

Topics:

- [Summary](#)

Summary

Boomi DCP gives organizations a framework and tools to reduce the complexity of large multisource data environments. The DCP framework enables IT to manage authentication and authorization for the data analyst community to protect the privacy of personally identifiable information (PII), and gives proper access to the proper teams and individuals. The multilevel, role-based access control (RBAC) provides sufficient flexibility while remaining consistent and manageable for even the largest organization. The use case scenario in this document demonstrates one example of how the DCP RBAC could enable three types of financial organization users to have the minimum data access rights, while still performing their respective job duties.

Boomi DCP can be added to your new or existing cluster as easily as adding a node to Hadoop. Using the integrated Cloudera wizard to extend a few Hadoop services to the Boomi node is all that is required to prepare to install the platform. DCP comes with its own installation scripts that are used once the cluster node is prepared.

Once a user has created a data set or been given access to an existing data set, they must understand the data size, distributions, and quality. The Dell Technologies use case showed how Boomi DCP provides easy-to-access and interpret summaries of the data set size, and a summary of each variable. The time savings that result from the platform generating this metadata is significant, even for small teams. The savings are even more substantial as the size of the supported analytics team expands. Centralizing data summarization also helps eliminate the potential for inconsistent conclusions. Such inconsistencies can occur when each analyst or team chooses their own tools and methods for data exploration. Any organization can benefit from spending less hours developing duplicate exploratory data analysis code and notebooks. Boomi DCP incorporates centralizing both externally generated and platform-generated metadata, using a single framework.

As organizations have brought more digital data under management, an unanticipated consequence is that fewer business analysts have the skills to work with these "big data" repositories. The professional data engineers and analysts can invest in "code first" approaches to working with big data. These approaches are largely focused on the Hadoop and NoSQL open-source tool sets. Boomi DCP also provides "low code-no code" capabilities that are accessible to the general business analyst community. These analysts have important roles that preclude investing in professional software development skills.

The Dell Technologies use case demonstrates how someone with a basic understanding of data structures and table joining can define a complete, end-to-end transformation using UI-based wizards and minimal coding. One of the more significant Boomi DCP differentiators is that the output of a data transformation job definition is highly efficient SQL code. That code can be immediately run or scheduled for later execution, on a fully featured, highly scalable Spark cluster. Boomi DCP re-engages the business analytics community that was primarily using spreadsheets or single-threaded Python or R on workstations, to generate value from the massive big data stores now available without learning scale-out Spark coding.

This document has also demonstrated the power of Boomi DCP to achieve additional value from existing or new Hadoop investments. The DCP framework enables IT to focus on the management of business-critical Hadoop investments, while supporting the needs of the business analyst community for accessing big data processing resources.

References

Additional information can be obtained at the [Dell Technologies Solutions InfoHub for Data Analytics](#). If you need additional services or implementation help, contact your Dell Technologies sales representative.

Topics:

- [Dell Technologies documentation](#)
- [Boomi documentation](#)
- [Cloudera documentation](#)
- [Dell Technologies InfoHub](#)
- [More information](#)

Dell Technologies documentation

The following Dell Technologies documentation provides additional and relevant information. Access to these documents depends on your login credentials. If you do not have access to a document, contact your Dell Technologies sales representative.

Table 7. Dell Technologies documentation

Document type	Location
Hadoop Infrastructure	Dell Technologies Cloudera CDP Data Center on Dell EMC Infrastructure Reference Architecture

Boomi documentation

The following documentation on the [Boomi website](#) provides additional and relevant information:

Table 8. Boomi documentation

Document type	Location
Boomi DCP	https://boomi.com/platform/data-catalog-and-preparation/
Boomi Documentation Portal	https://help.boomi.com/
Boomi DCP Connectors	https://boomi.com/platform/data-catalog-and-preparation/connectors/

Cloudera documentation

The following documentation on the [Cloudera documentation website](#) provides additional and relevant information:

Table 9. Cloudera documentation

Document type	Location
Cloudera Data Hub	https://www.cloudera.com/products/data-hub.html
Cloudera Support Portal	https://my.cloudera.com/support.html

Table 9. Cloudera documentation (continued)

Document type	Location
CDP Private Cloud Base Installation Guide	https://docs.cloudera.com/cdp-private-cloud-base/7.1.3/installation/topics/cdpdc-installation.html
CDP Private Cloud Base Production Installation	https://docs.cloudera.com/cloudera-manager/7.1.1/installation/topics/cdpdc-prod-installation.html

Dell Technologies InfoHub

The [Dell Technologies InfoHub](#) is your one-stop destination for the latest information about Dell EMC Solutions and Networking products. New material is frequently added, so visit often to keep up to date on our expanding portfolio of cutting-edge products and solutions.

More information

For more information, contact your Dell Technologies or authorized partner sales representative.