

Auslagern von Data Warehousing nach Hadoop mithilfe von DataStage

Verwenden Sie IBM® InfoSphere® DataStage®, um Hadoop zu laden und YARN zum Verwalten von DataStage-Workloads in einem Hadoop-Cluster zu verwenden

Diese Umgebung ist eine Front-End-Windows-Workstation, auf der die DataStage Designer Version 11.5 installiert ist. Es stellt eine Verbindung zu einem Back-End-Linux-Server her, auf dem die Information-Server Server-Komponenten ausgeführt werden, und einer Konfiguration, die als Big-Integrate bezeichnet wird. Die DataStage Engine wird in einem remoten Hadoop-Cluster ausgeführt und nimmt an der Ressourcenverwaltung des Yarn-Dienstes in Hadoop teil.

- Vorkonfigurierte, automatisch gestartete Software und Dienste
- Ermöglicht es Ihnen, die Konfiguration zu sehen, die erforderlich ist, um den Hadoop-Cluster für die Ausführung von ETL-Prozessen zu verwenden

Tutorial

In dieser Demo verwenden Sie DataStage, um die Datenverarbeitung zum Extrahieren, Transformieren und Laden (ETL) in einem herkömmlichen Enterprise Data Warehouse abzuschließen. Anschließend werden die Daten- und ETL-Verarbeitung in skalierbare, hochwertige Hadoop-Cluster und Data Lakes ausgeladen.

In dieser Produkttour erhalten Sie Erfahrungen mit den folgenden Funktionen:

- Erfahren Sie, wie Sie traditionelle DataStage-ETL-Aufträge ausführen
- Konfigurieren von DataStage für die Ausführung in Hadoop-Clustern
- Überprüfen von Ausführungsprotokollen, um die Konfiguration ordnungsgemäß zu erstellen

Folgen Sie den Anweisungen in diesem Bereich, um die Demo im linken Bereich zu durchlaufen.

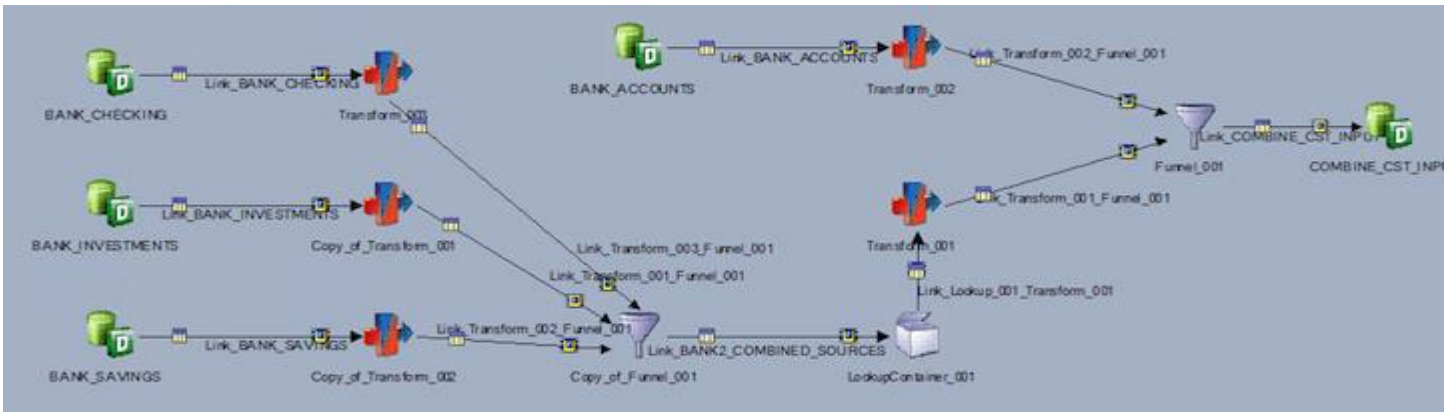
Ausführen eines herkömmlichen Data Warehouse-ETL-Auftrags

Zunächst überprüfen Sie einen DataStage-Auftrag, der Daten aus zwei Geschäftsbereichen in einer Tabelle kombiniert. Anschließend führen Sie den Auftrag aus, um ein kombiniertes Datenrepository zu erstellen.

1. Klicken Sie jetzt auf **Demo starten**
2. Klicken Sie im Startmenü auf das Symbol IBM InfoSphere DataStage und QualityStage Designer.



Der Auftrag DS07-JK-BANK1-UND-JK-BANK2-To-COMBINE-CST-INPUT wird angezeigt.

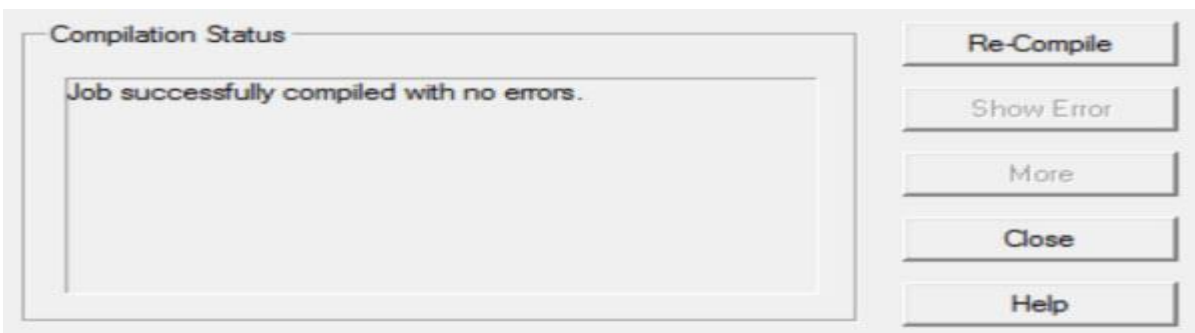
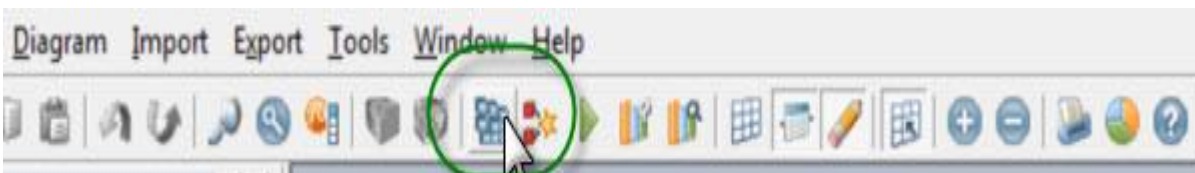


Der Auftrag hat die folgenden Datenquellen:

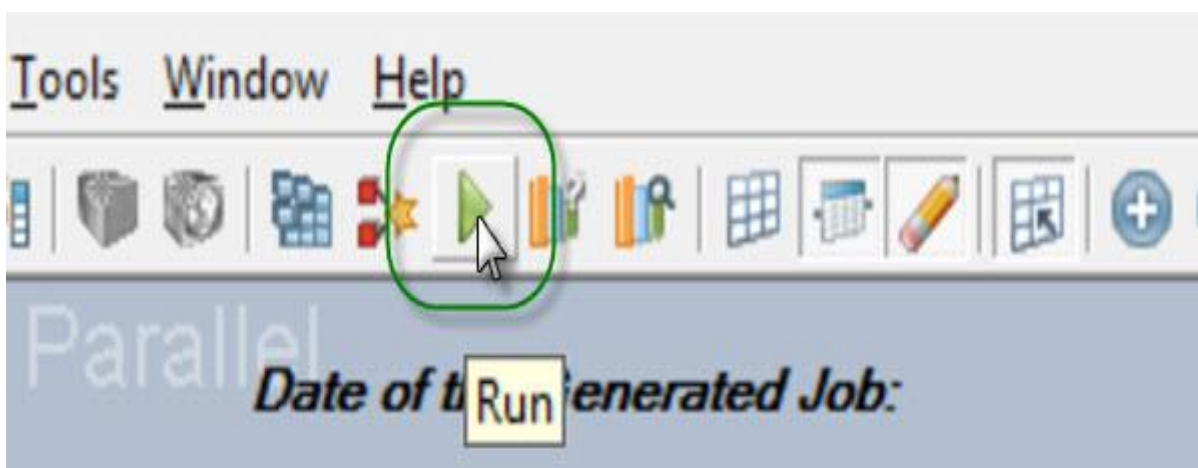
- * BANK1 ACCOUNTS-Daten, die als "BANK_ACCOUNTS" gekennzeichnet sind.
- * BANK2 CHECKING Kundendaten, die als "BANK_CHECKING" gekennzeichnet sind.
- * BANK2 INVESTMENT Kundendaten, die als "BANK_INVESTMENTS" gekennzeichnet sind.
- * BANK2 SAVINGS Kundendaten, die als "BANK_SAVINGS" gekennzeichnet sind.

Das Datenziel des Auftrags sind BANK1 und BANK2 kombinierte Kundendaten.

3. Kompilieren Sie den Auftrag, indem Sie auf das **Symbol Kompilieren** auf der Symbolleiste klicken.



4. Führen Sie den Auftrag aus, indem Sie auf das Symbol **Ausführen** auf der Symbolleiste klicken.



5. Im Fenster Auftragsausführungsoptionen können Sie Werte für die Laufzeitauftragsparameter angeben. Verwenden Sie die Standardwerte. Geben Sie für den Parameter 'JKLW_DBS_PWD' 'inf0server' ein. Klicken Sie auf **Ausführen**, um den Auftrag auszuführen.

6. Überprüfen Sie nach dem Ausgeführt des Auftrags die Daten aus der Ausgabephase:

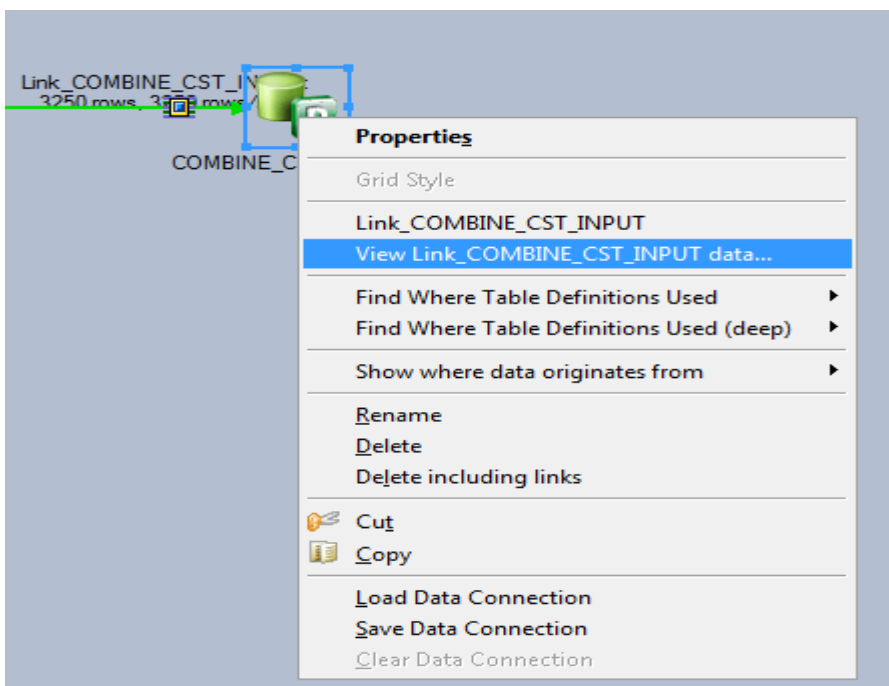
Name	Value
JKLW_DBS_PWD	
APT_YARN_CONF	/opt/IBM/InformationServer/Server/PXEngine/etc/yam_conf/yamconfig.cf
Configuration file	/opt/IBM/InformationServer/Server/Configurations/default.ap
parameters paramet	(As pre-defined)
hostname	bigdata.ibm.com
Use IPv4 only	True

Workload management

Queue: Project default (HighPriorityJobs)

Buttons: Run, Validate, Cancel, Help

Eine. Klicken Sie auf den **Combine-CST-INPUT-Ausgabedatenbankconnector**, und klicken Sie auf Link anzeigen, COMBINE-CST-INPUT-Daten



Die zurückgegebenen Daten stellen einen kombinierten Kundeneingabesatz für JKBANK1 & JK-BANK2-Unternehmenskunden in nicht standardisierter und nicht bereinigter Form dar.

ACCOUNT_HOL...	ACCOUNT_ID	ACCOUNT_TYPE	ADDR1	ADDR2	BANKCARD
	30103176	INVESTMENT	9600 SE 36th Ave		YES
	60103134	INVESTMENT	10329 E 61st St		YES
	40103052	INVESTMENT	9633 W Ferris Branch Blvd		YES
	30102326	INVESTMENT	535 S Denby Ave		YES
	40102947	INVESTMENT	213 Maple St		YES
	40102482	INVESTMENT	11155 Ruesta Dr		YES
	20102480	INVESTMENT	7347 Lobdell Rd		YES
	30102386	INVESTMENT	14401 Inkster Rd		YES
	30102531	INVESTMENT	11002 Plumewood Dr		YES
	40102662	INVESTMENT	567 Amberina Ct		YES
	60102594	INVESTMENT	Route 15		YES
	20102995	INVESTMENT	8513 Peachwood Dr		YES
	50102833	INVESTMENT	52 Main Dr		YES
	30102491	INVESTMENT	206 Albright Rd		YES
	40103137	INVESTMENT	700 Veterans Dr		YES
	40101702	SAVINGS	37816-a 44th Ave		NO
	60101634	CHECKING	1708 N Damen Ave		NO

B. Klicken Sie auf **Schließen**, um das Fenster zu schließen.

Verschieben von Daten aus Data Warehouse nach Hadoop

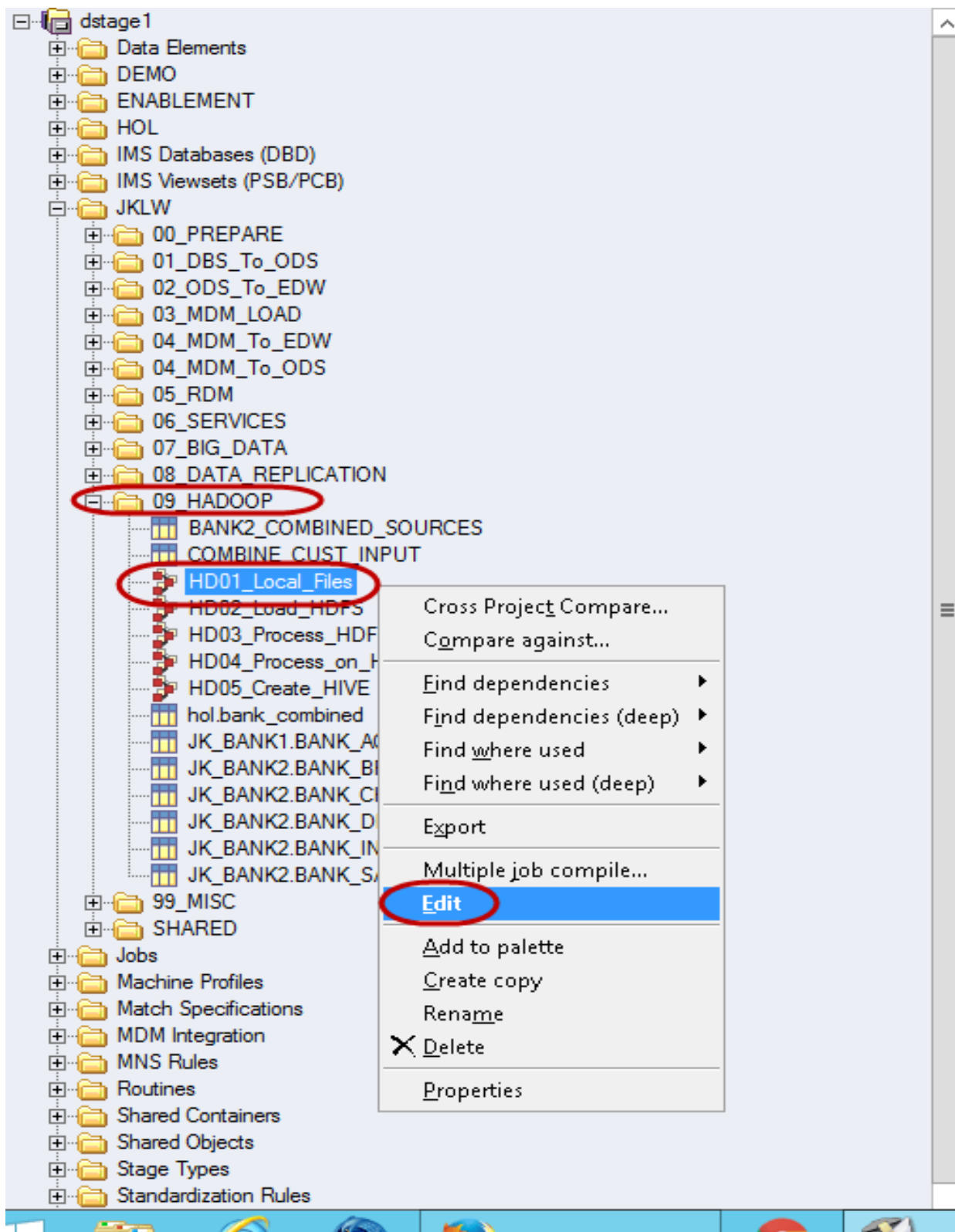
Erfahren Sie, wie Sie die Skalierbarkeit des parallelen Moduls nutzen und die Verarbeitung aus dem Enterprise Data Warehouse in einen Hadoop-Cluster auslagern.

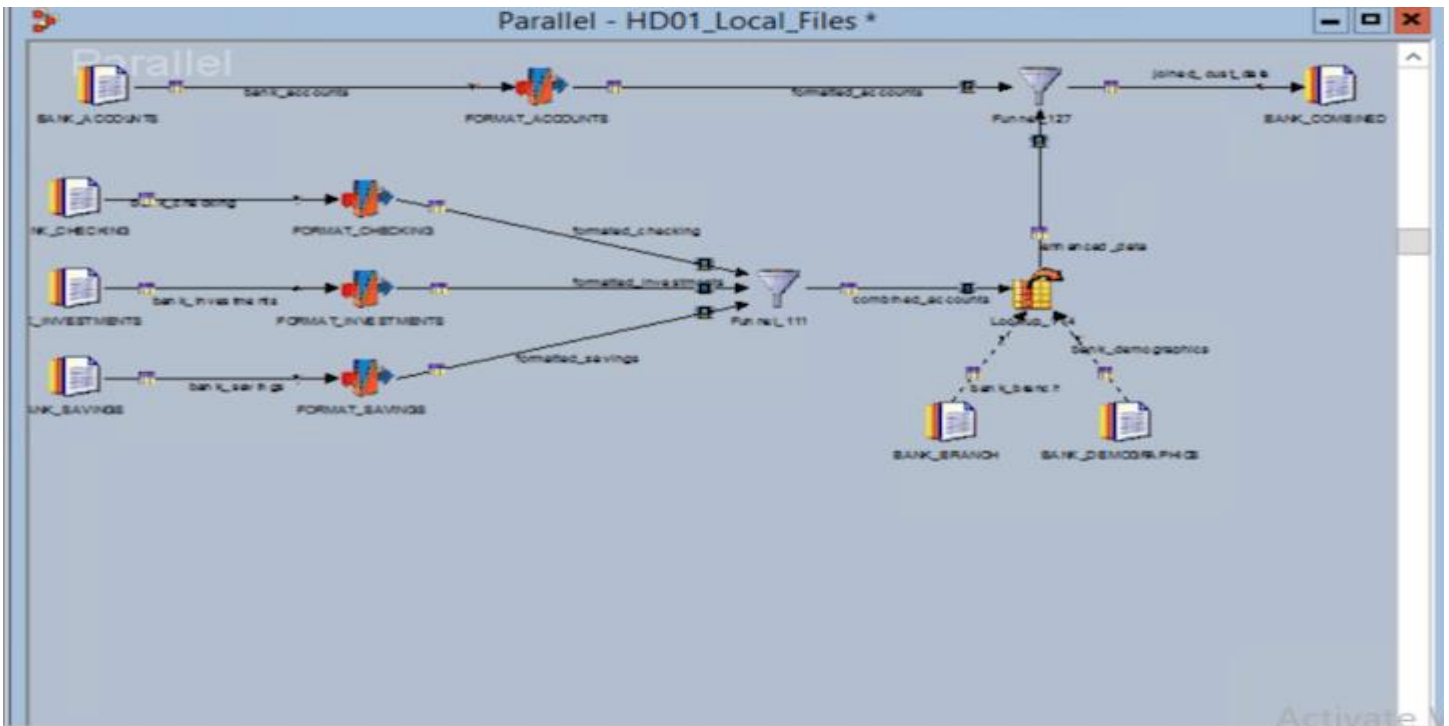
In dieser Aufgabe führen Sie die folgenden Schritte aus:

- Ausführen eines Dateiverarbeitungsauftrags nativ auf dem Linux-Hostsystem
- Push the data to the Hadoop cluster

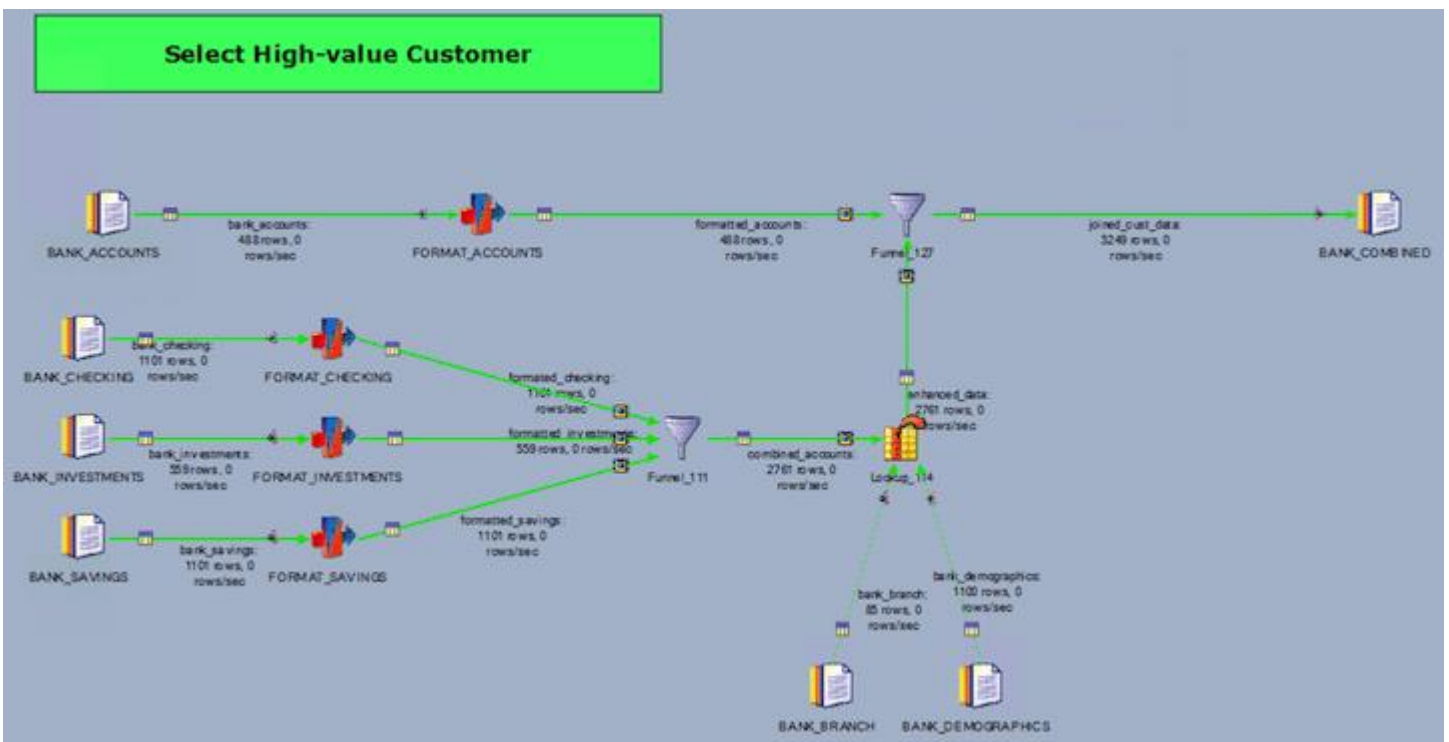
In diesem Auftrag werden Dateien, die auf Linux basieren, auf dem DataStage-Leiterknoten als Landezone aufgenommen. Nachdem die Dateien aufgenommen wurden, verarbeitet DataStage die Datendateien in einer herkömmlichen DataStage-Konfiguration.

1. Klicken Sie auf den Job **HD01-Lokal/Dateien**, und klicken Sie auf **Bearbeiten**.

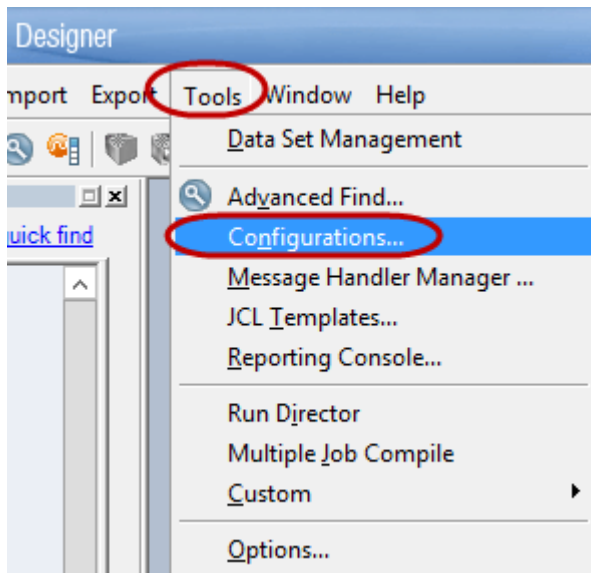




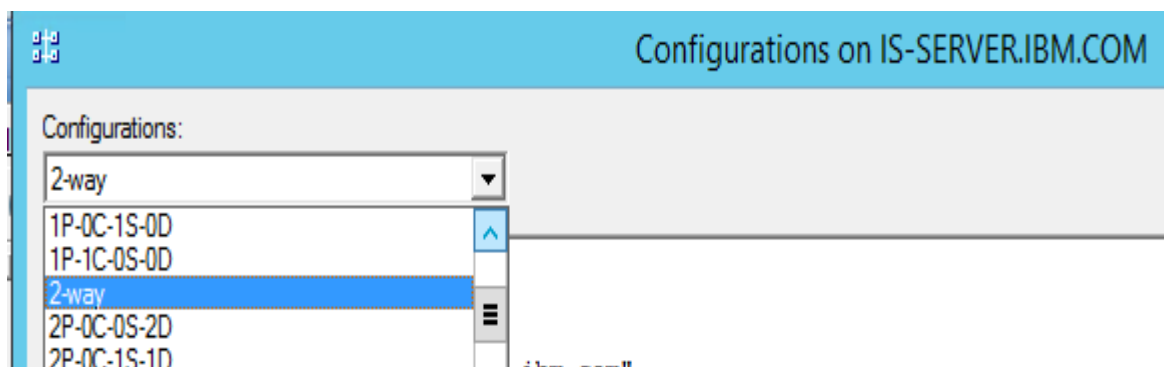
2. Compile the job by clicking the **Compile** icon on the toolbar.
 3. Run the job by clicking the **Run** icon on the toolbar.
 4. In the Job Run Options window, you can select the configuration file. For this demo, use the `default.apr` configuration file. Click **Run**.
- When the job is finished, all the job links turn green and show the number of rows on each link.



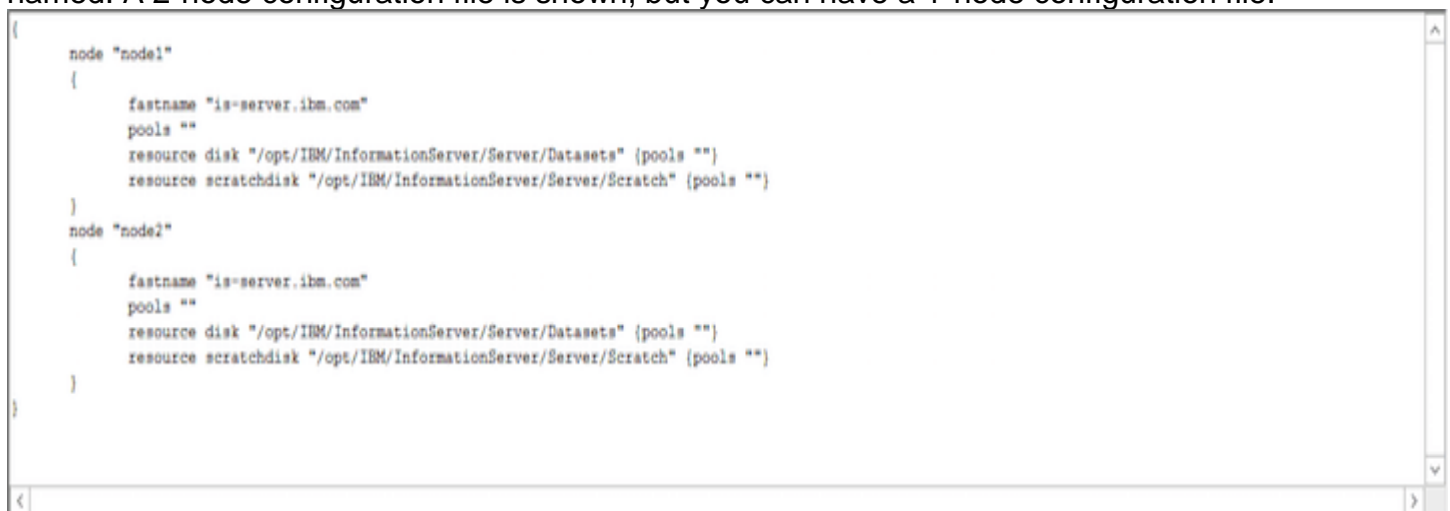
5. Review the default configuration file:
 - a. Click **Tools & Configurations**.



b. From the Configurations list, select 2-way.



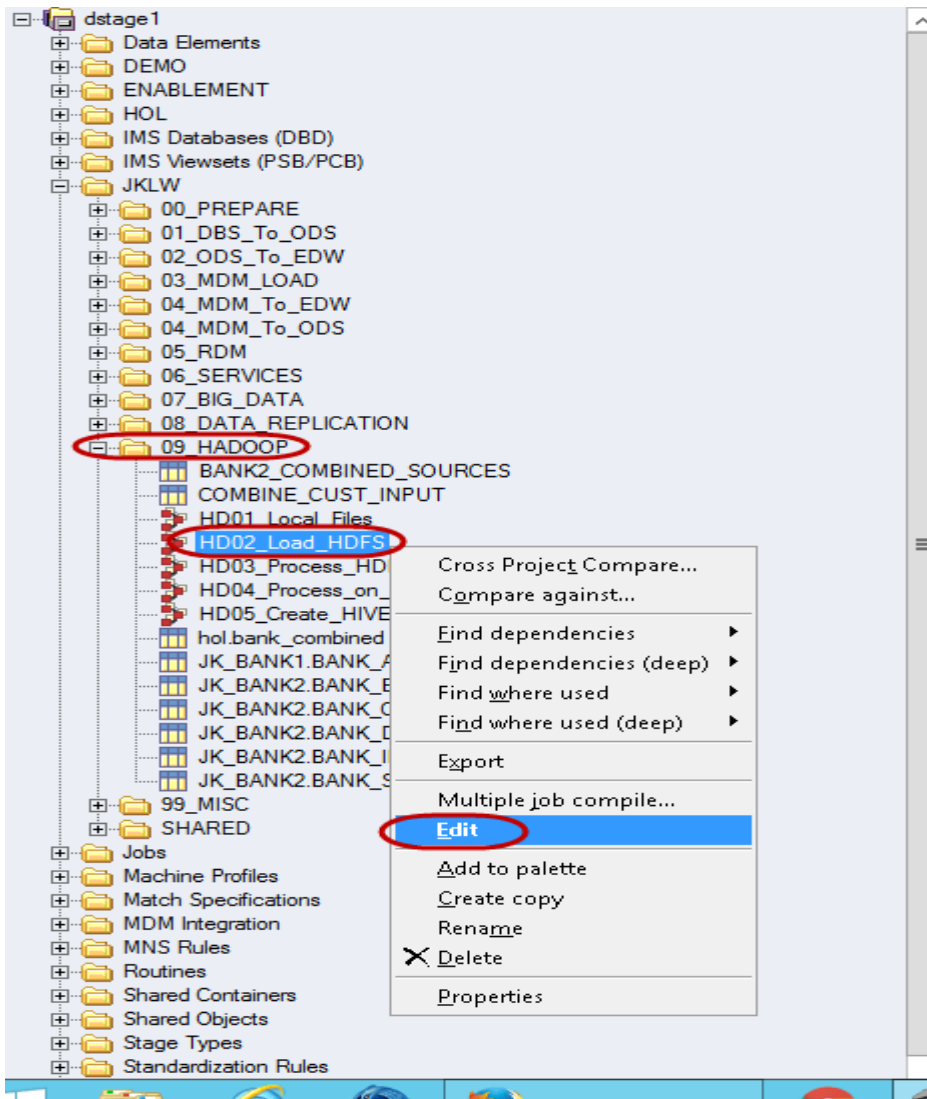
This file is a traditional configuration file, where the nodes and the location of the resources are named. A 2-node configuration file is shown, but you can have a 1-node configuration file.



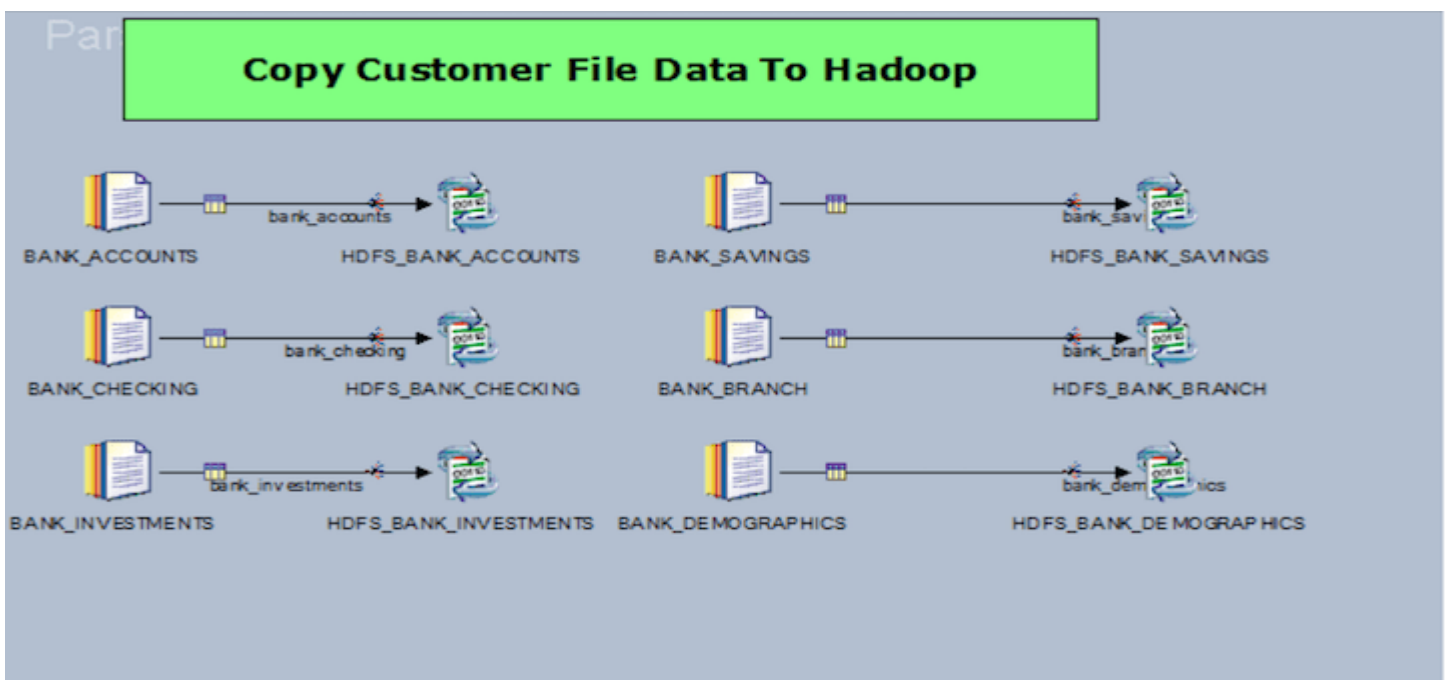
c. Click Close.

Load data into the Hadoop HDFS

1. Click the **HD02Load\HDFS** job and click **Edit**.

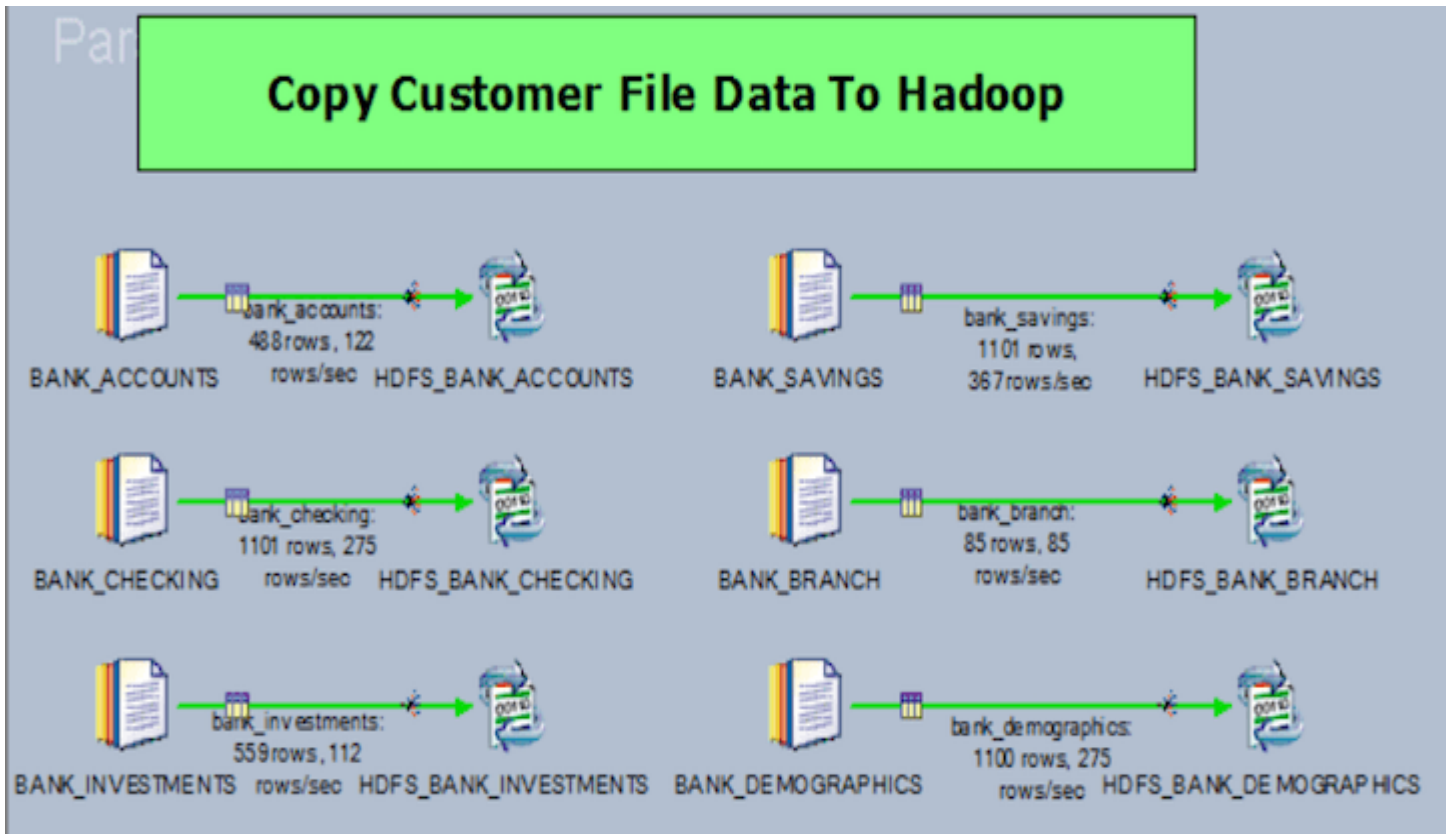


The HD02_Load_HDFS job loads your files onto HDFS for processing later.



2. Compile the job by clicking the Compile icon on the toolbar.
3. Run the job by clicking the Run icon on the toolbar.
4. In the Job Run Options window, you can select the configuration file. Use the `default.apf` configuration file. Click **Run**.

When the job is finished, all the job links turn green and show the number of rows on each link.



Überprüfen einer File Connector-Eigenschaft

1. Klicken Sie auf den Dateiconnector **HDFS-BANK-ACCOUNTS**.

Das Fenster Dateiconnector zeigt, wie Sie die Konnektivitätsattribute für den Big Data-Server zugewiesen haben. Sie können auch die Dateiattribute anzeigen, wo und wie Sie die Daten schreiben. Der Host- und Dateipfad kann unterschiedlich sein.

Stage | Input |

Stage name
HDFS_BANK_ACCOUNTS

General | Properties | Advanced |

File system	WebHDFS
Use custom URL	No
Use SSL (HTTPS)	No
Use Kerberos	No
Use keytab	No
Host *	is-server.ibm.com
Port	
Service principal	
User name *	dsadm
Password	
Keytab *	
Custom URL *	

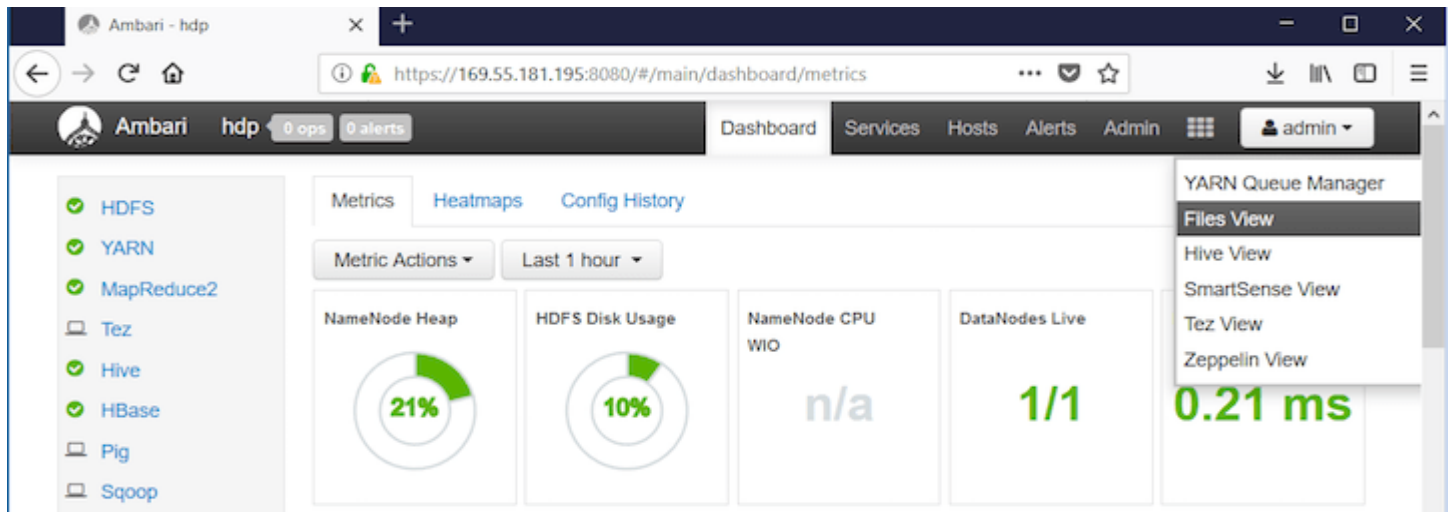
▼ Usage

Write mode	Write single file
File name *	ikw/HDFS_BANK_ACCOUNTS.TXT
If file exists	Overwrite file
▶ Split file on key changes	No
Maximum file size	0
Force sequential	No
Cleanup on failure	Yes
▼ File format	Comma-separated value (CSV)
▶ Avro format properties	
▶ ORC Settings	
▼ Delimited format properties	
▼ Encoding	
Include byte order mark	No
▼ First row is header	Yes
Include data types	Yes
Field delimiter	,
Row delimiter	<NL>
Escape character	
Quotation mark	None
Null value	""
▼ Field format	

OK Cancel Help

Wenn Sie die Überprüfung der Informationen abgeschlossen haben, klicken Sie auf **OK**, um das Eigenschaftfenster zu beenden.

2. Überprüfen Sie die Daten im HDFS-Dateisystem. Die Ambari Webkonsole ist geöffnet und der Benutzeradministrator ist eingeloggt. 3. Klicken Sie in der Ambari-Konsole auf **Dateiansicht**.



4. Klicken Sie in der Liste der Verzeichnisse auf **Benutzer**.







Name	Size	Last Modified	Owner	Group
app-logs	--	2017-09-07 19:27	yarn	hadoop
apps	--	2017-06-26 00:40	hdfs	hdfs
ats	--	2017-06-25 07:23	yarn	hadoop
biginsights	--	2017-06-26 01:08	hdfs	hdfs
hdp	--	2017-06-25 07:23	hdfs	hdfs
mapred	--	2017-06-25 07:23	mapred	hdfs
mr-history	--	2017-06-25 07:23	mapred	hadoop
spark-history	--	2017-08-04 15:49	spark	hadoop
spark2-history	--	2018-03-26 15:22	spark	hadoop
tmp	--	2018-03-24 15:19	hdfs	hdfs
user	--	2017-08-18 10:16	hdfs	hdfs

5. Klicken Sie auf das **Verzeichnis dsadm**.

ambari-qa	--	2017-08-04 13:39	ambari-qa	hdfs
as_user	--	2017-07-04 17:00	as_user	hdfs
biadmin	--	2017-08-18 09:54	admin	hdfs
bigsql	--	2017-06-26 17:41	bigsql	hdfs
dsadm	--	2018-03-08 22:19	dsadm	dstage
hbase	--	2017-06-25 07:23	hbase	hdfs
hcat	--	2017-06-26 00:40	hcat	hdfs
hive	--	2017-08-04 12:51	hive	hdfs


6. Klicken Sie auf den **Ordner jklw**.



Die Dateien, die Sie in das HDFS-Dateisystem geladen haben, befinden sich im Ordner.









 hdfs1_yarn.txt	0.1 kB	2018-03-08 22:19	hdfs	dstage
 hdfs1_yarn_debug.txt	0.1 kB	2017-09-28 13:01	hdfs	dstage
 insurance	--	2018-02-28 11:26	dsadm	dstage
 jklw	--	2018-03-26 12:48	dsadm	dstage
 my_hdfs_yarn_txt	693.9 kB	2018-03-08 22:19	hdfs	dstage
 sample	--	2018-03-07 19:01	dsadm	dstage



Für die Zwecke dieser Demo, siehst du die HDFS_BANK_ACCOUNTS.TXT-Datei.

7. Klicken Sie in der oberen Menüleiste auf **Öffnen**.

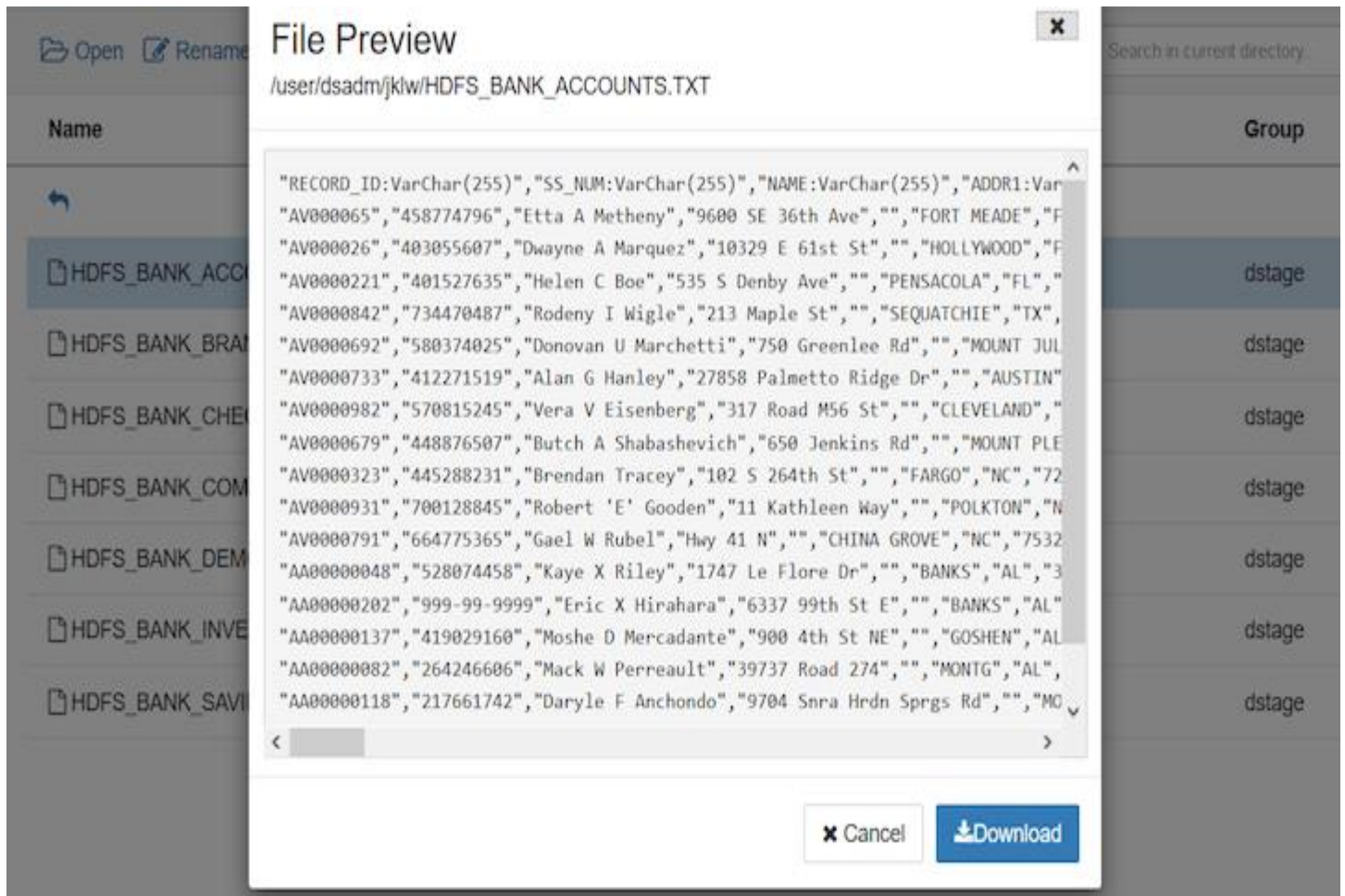
 Ambari hdp 0 ops 0 alerts Dashboard Services Hosts Alerts Adm

  / > user > dsadm > jklw 1 Files, 0 Folders selected

 Open  Rename  Permissions  Delete  Copy  Move  Download  concatenate

Name	Size	Last Modified	Owner
 HDFS_BANK_ACCOUNTS.TXT	117.4 kB	2018-03-28 20:04	dsadm
 HDFS_BANK_BRANCH.TXT	4.2 kB	2018-03-28 20:04	dsadm

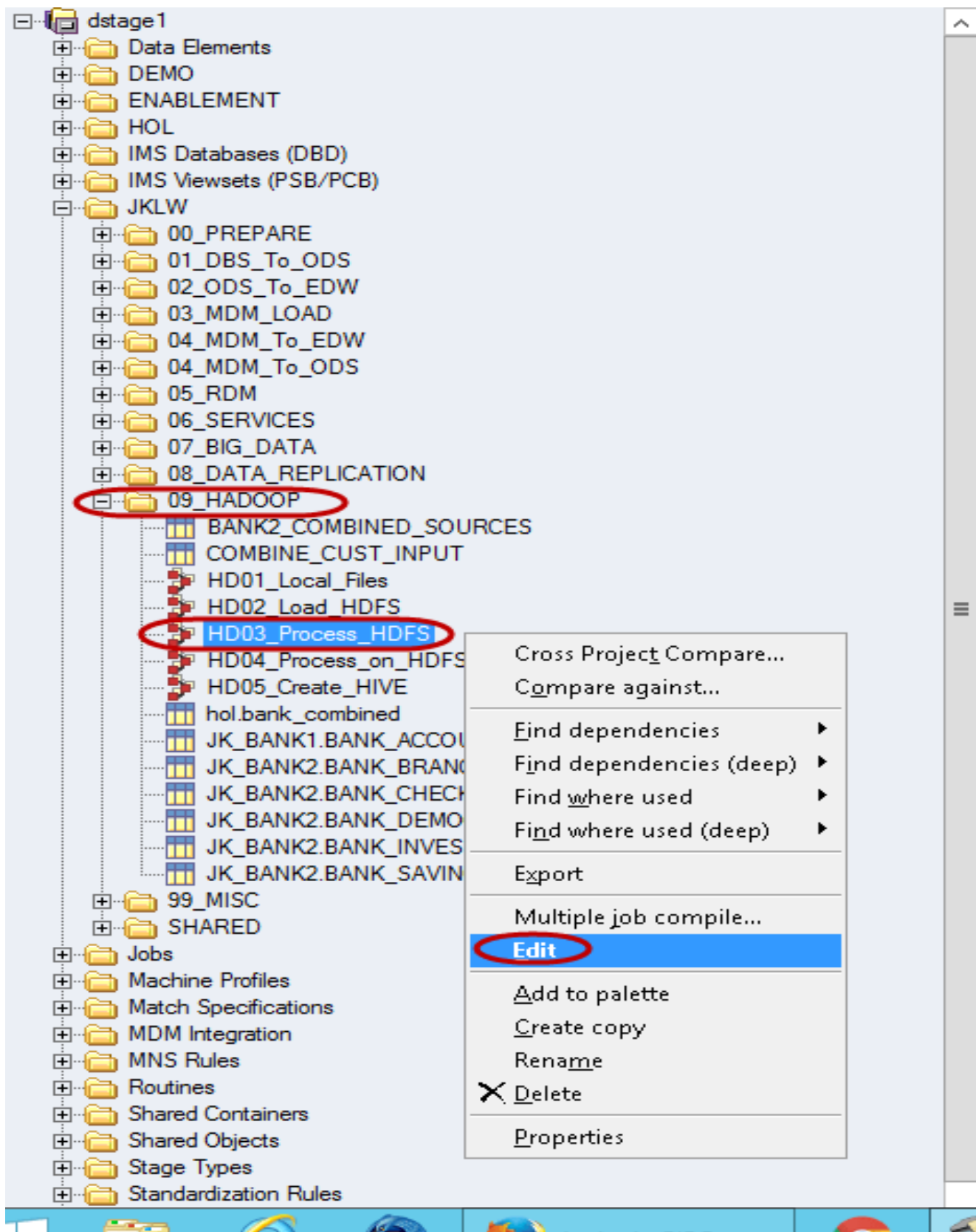
8. Überprüfen Sie die Daten in der Datei. Wenn Sie fertig sind, klicken Sie auf **Abbrechen**.



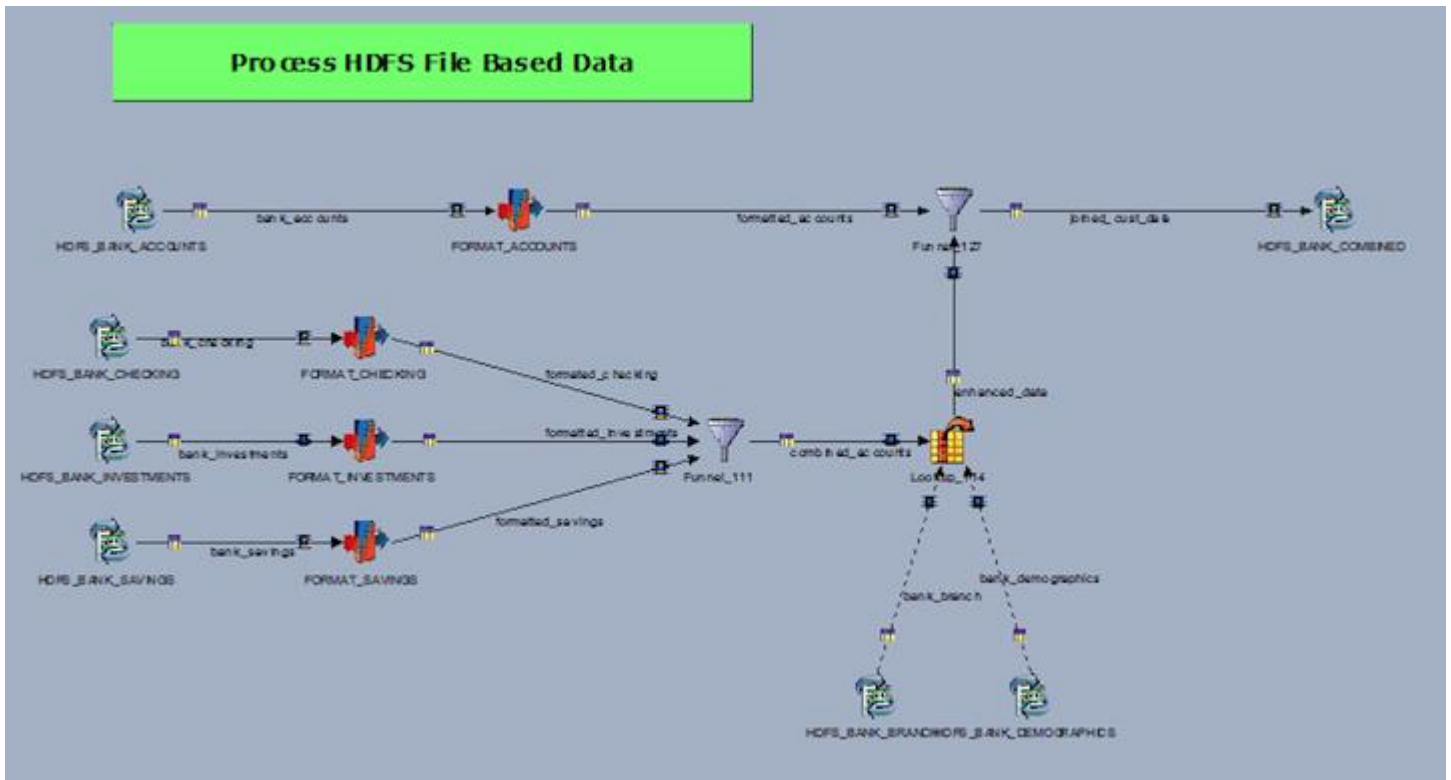
Verarbeiten der Hadoop-Dateien

Verarbeiten Sie die HDFS-dateibasierten Daten in einer herkömmlichen DataStage-Konfiguration.

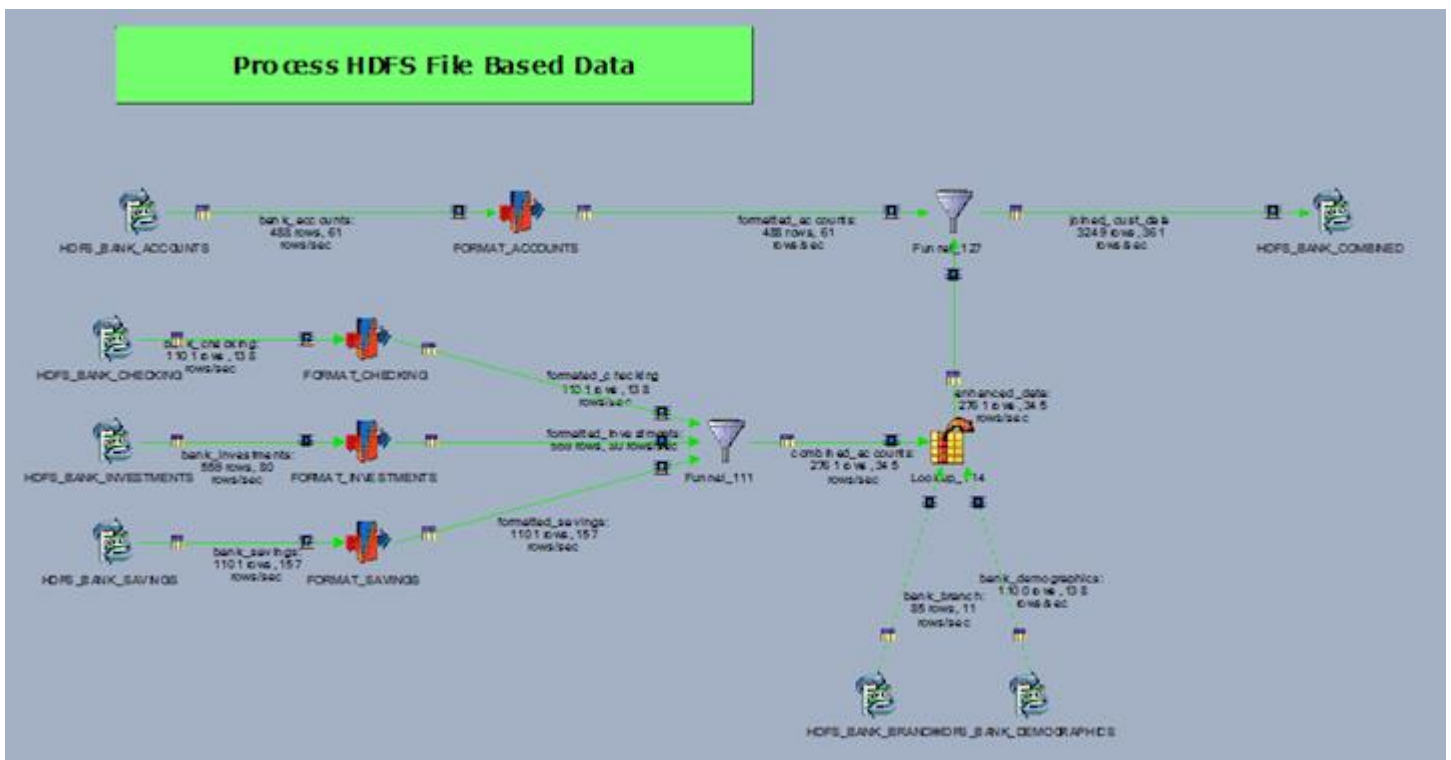
1. Klicken Sie auf den Job **HD03-Prozess-HDFS** und dann auf Bearbeiten.



Der HD03-Load-HDFS-Auftrag liest Daten aus Ihrem Hadoop HDFS-Dateisystem, verarbeitet die Daten und schreibt die konsolidierten Kundendaten zurück in das Hadoop HDFS-Dateisystem.










2. Kompilieren Sie den Auftrag, indem Sie auf das Symbol **Kompilieren** auf der Symbolleiste klicken.
 3. Führen Sie den Auftrag aus, indem Sie auf das Symbol **Ausführen** auf der Symbolleiste klicken.
 4. Im Fenster Job Run Options können Sie die Konfigurationsdatei auswählen. Verwenden Sie die Konfigurationsdatei 'default.apt'. Klicken Sie auf **Ausführen**.
- Wenn der Auftrag abgeschlossen ist, werden alle Auftragsverknüpfungen grün und zeigen die Anzahl der Zeilen auf jedem Link an.


















5. Return to the HDFS File view by clicking the Maximize Window icon.

Notice the time that the HDFS\BANK\COMBINED\TEXT file was written onto the Hadoop file system.

 HDFS_BANK_ACCOUNTS.TXT	117.4 kB	2018-03-28 20:04
 HDFS_BANK_BRANCH.TXT	4.2 kB	2018-03-28 20:04
 HDFS_BANK_CHECKING.TXT	174.3 kB	2018-03-28 20:04
 HDFS_BANK_COMBINED.TXT	832.6 kB	2018-03-28 13:06
 HDFS_BANK_DEMOGRAPHICS.TXT	137.0 kB	2018-03-28 20:04
 HDFS_BANK_INVESTMENTS.TXT	83.6 kB	2018-03-28 20:04
 HDFS_BANK_SAVINGS.TXT	161.4 kB	2018-03-28 20:04

6. Click the **refresh icon** at the top of the File view to reload the file information.

 Open  Rename  Permissions  Delete  Copy  Move  Download 		
Name	Size	Last Modified
		
 HDFS_BANK_ACCOUNTS.TXT	117.4 kB	2018-03-28 20:04
 HDFS_BANK_BRANCH.TXT	4.2 kB	2018-03-28 20:04
 HDFS_BANK_CHECKING.TXT	174.3 kB	2018-03-28 20:04
 HDFS_BANK_COMBINED.TXT	832.6 kB	2018-03-28 20:15
 HDFS_BANK_DEMOGRAPHICS.TXT	137.0 kB	2018-03-28 20:04
 HDFS_BANK_INVESTMENTS.TXT	83.6 kB	2018-03-28 20:04

The file time stamp is updated after the job ran.

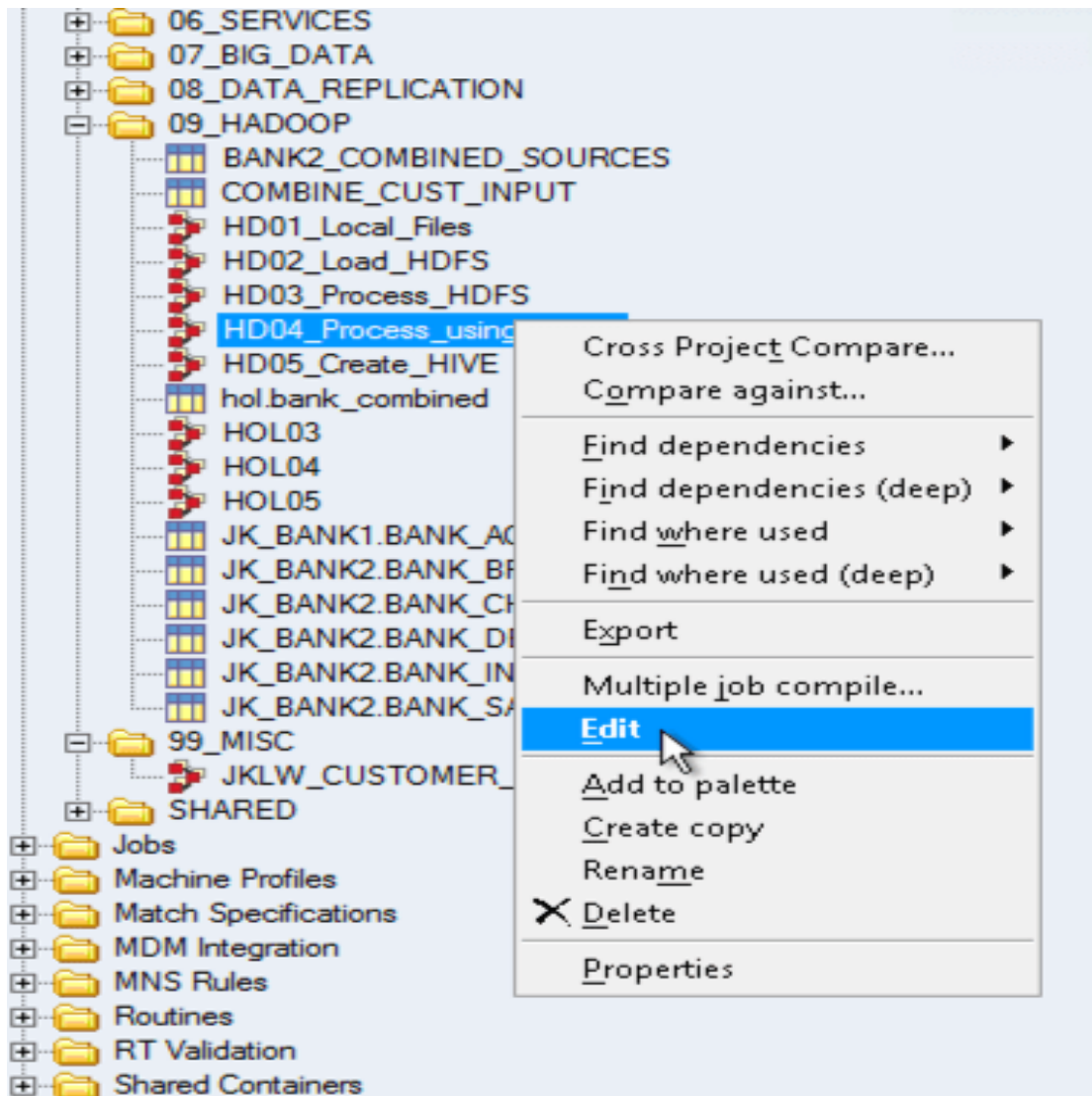
Close the File view by clicking the highlighted file.

7. Return to DataStage and close the job.

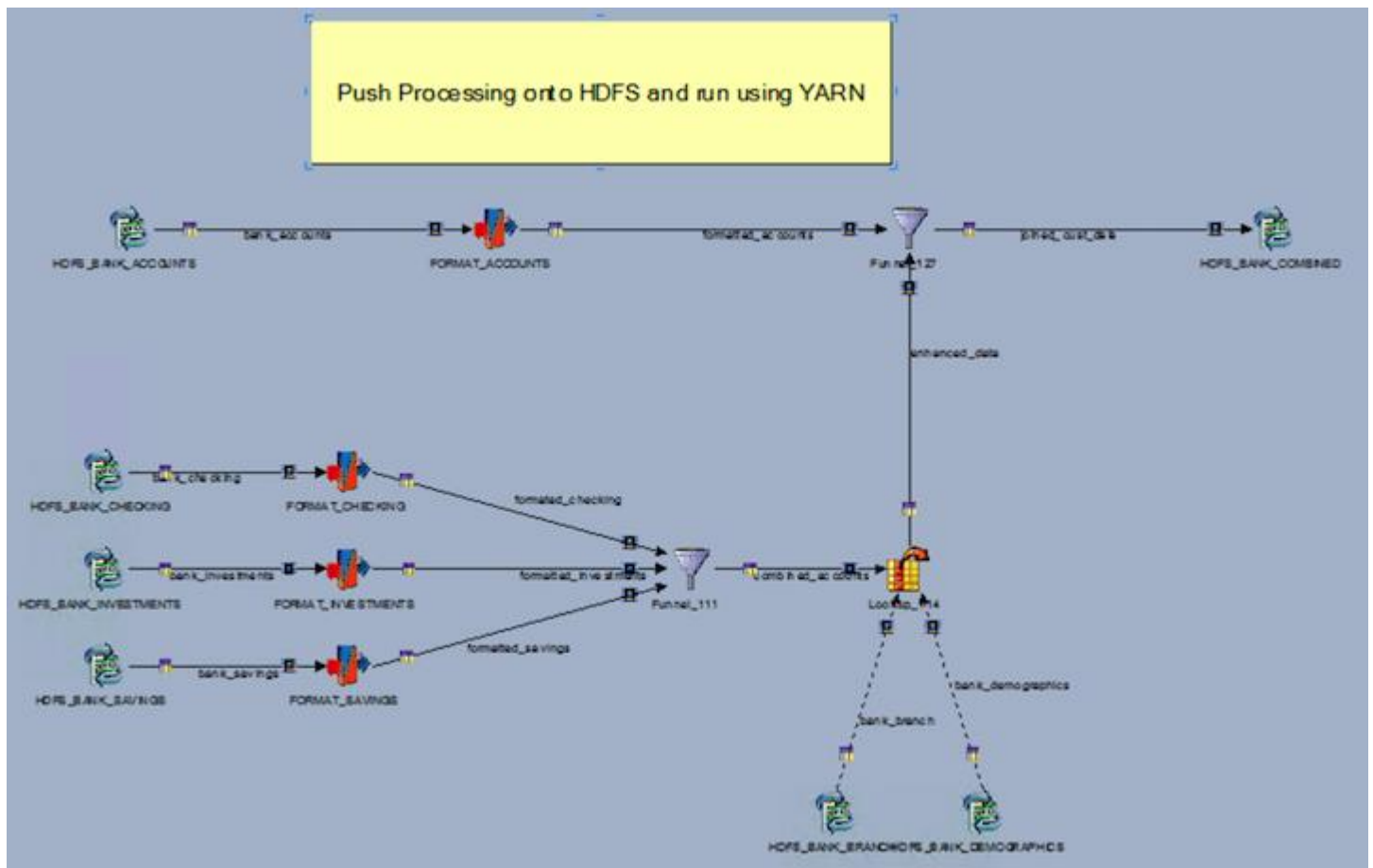
Run ETL processing inside Hadoop by using YARN

Process the HDFS-file-based data by pushing the processing to the Hadoop cluster and allowing the processing to run by using YARN.

1. Right-click the HD04\Process\using\YARN job and click **Edit**.



The HD04\Process\on\HDFS\using\YARN job reads data from your Hadoop HDFS file system, processes the data, and writes the consolidated customer data back to the Hadoop HDFS file system entirely in the Hadoop cluster.



2. Compile the job by clicking the **Compile** icon on the toolbar. Run the job by clicking the **Run** icon on the toolbar.

In the Job Run Options window, you can see that you have two properties.

The first property points to the `yarnconfig.cfg` file. This file indicates to DataStage which settings it needs to communicate with and run the process on YARN.

The second property points to the configuration file so that DataStage detects how many nodes to run and where the resources are.

3. Before you click **Run**, learn about the properties:

- * On the server, the `yarnconfig.cfg` file contains a number of settings to tune how DataStage processes are run with YARN. Two important settings are `APT_YARN_MODE` and `APT_YARN_USE_HDFS`.

- * The `APT_YARN_MODE` setting tells the engine where to run. A value of `false` tells the engine to run normally on the host system. A value of `true` tells the engine to hand the process to YARN for processing on the Hadoop cluster.

- * The `APT_YARN_USE_HDFS` setting tells the engine whether the data resources are being written to the local file system or the HDFS file system. A setting of `false` indicates that the data is written to the local file system. A setting of `true` indicates that the data is being written to the HDFS file system.

- * You can use many other settings to further tune the interaction. Each setting is described in the `yarnconfig.cfg` file.


```

## Licensed Materials - Property of IBM
## (c) Copyright IBM Corp. 2015
#
# DataStage PX Yarn Configuration
# =====
# Lines in this file are either comments, introduced by a # sign like this,
# or of the form "key=value". Key lines may be commented out below.
#
# IMPORTANT:
# Ensure when making changes to this file that it is saved with the encoding set to
# UTF-8. Please be aware if the encoding isn't set to UTF-8 this may produce undesired
# behaviour.

APT_YARN_MODE=true
# If defined and set to 1 or true runs the given PX job on
# the local Hadoop install in YARN mode.

APT_YARN_CONTAINER_VCORES=0
# Defines the number of virtual cores that the containers will request to run
# PX Section Leader and Player processes in.
# The default is 0 which means "Don't set it".

APT_YARN_CONTAINER_SIZE=64
# Defines the size in MBs of the containers that will be requested to run
# PX Section Leader and Player processes in.
# The default is 64MB if not set.

APT_YARN_CONTAINER_SIZE_AUTO=false
# When defined will automatically use the estimated container size for the largest partition
# as the container size if it is larger than the set container size.
# It accepts a value of true or false.

APT_YARN_BASE_PROCESS_SIZE=16
# Defines the base process size in MB to be used when estimating the size of containers
# to request from YARN. The default value for this is 16MB.

APT_YARN_ALLOCATION_TIMEOUT=180
# Specifies the amount of time in seconds to wait for allocations of containers

```

* When parallel jobs run on Hadoop, the jobs request a set of containers from YARN. The containers represent the resources that the job was allocated. Each resource has a designated amount of virtual CPU and memory for each container. The number of containers that are requested is equal to the number of logical nodes that are defined in the `APT_CONFIG_FILE` file.

* These files can be configured in one of three ways: static, dynamic, and mixed.

* A static file looks like a regular `APT_CONFIG_FILE`, except the resource disk can be either on the local file system or in HDFS depending on the `APT_YARN_USE_HDFS` setting in the `yarnconfig.cfg` file.

* A dynamic configuration file uses the same format as the static Information Server configuration files, which assign fixed nodes to the job. However, a dynamic configuration file uses a fastname value of `\$\$host`, as opposed to a static configuration file that usually contains a host name. One node in the configuration file must contain the engine tier node, but this node can be defined with a conductor node pool if you don't want to run data processing on the conductor node.

* A mixed configuration file contains a mix of static host names (an actual host name) and dynamic host names (fastname `\$\$host`). The following configuration file specifies that the first 30 nodes

are defined by YARN. Then, the next 10 nodes are run on machineA, and the 41st node runs on machineB:

```
{
  node "node1"
  {
    fastname "mymachine.domain.com"
    pools ""
    resource disk "/mydisk/tmp" {pools ""}
    resource scratchdisk "/myscratch/tmp" {pools ""}
  }
}
```

```
{ node "node0"
{
  fastname "the-engine-tier-machine.domain.com"
  pools "conductor"
  resource disk "/sandbox/bsmith/tmp" {pools ""}
  resource scratchdisk "/scratch" {}
}
  node "node1"
  {
    fastname "$host"
    pools ""
    resource disk "/sandbox/bsmith/tmp" {pool ""}
    resource scratchdisk "/sandbox/bsmith/tmp" {pools ""}
    instances 30
  }
}
```



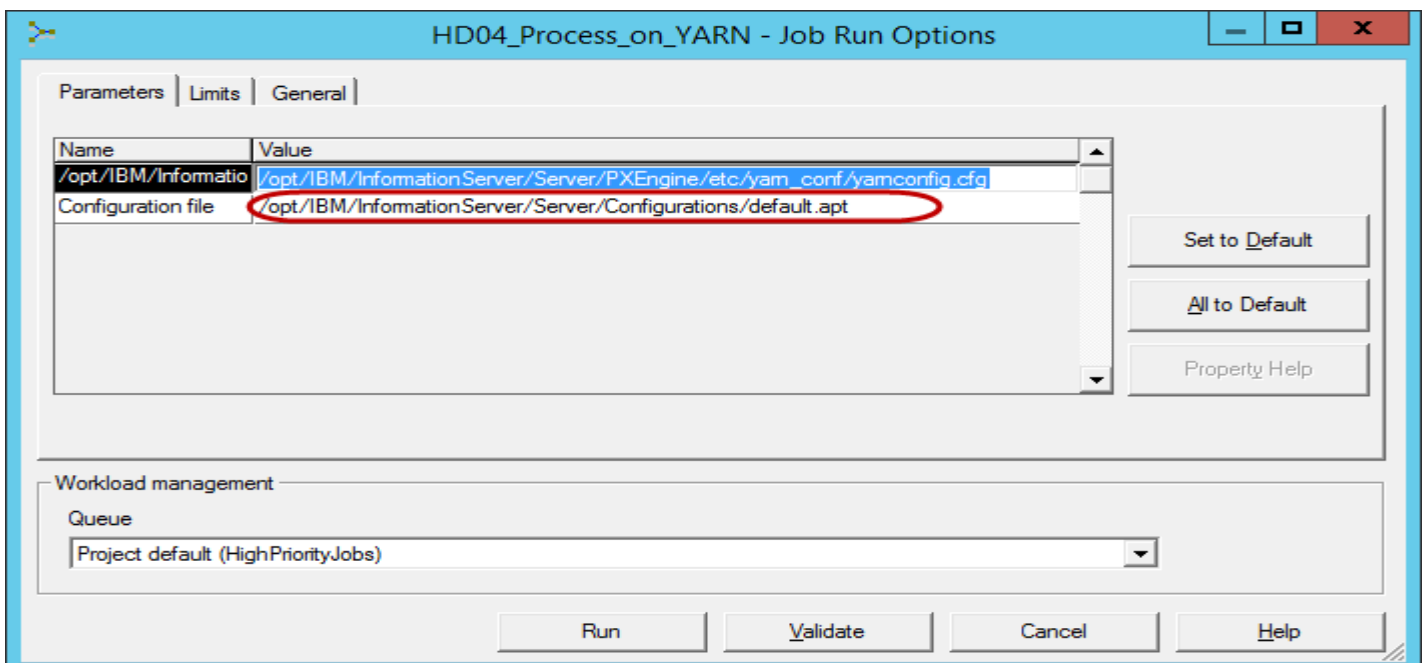
```

{
  node "node0"
  {
    fastname "the-engine-tier-machine.domain.com"
    pools "conductor"
    resource disk "/sandbox/bsmith/tmp" {pools ""}
    resource scratchdisk "/scratch" {}
  }
  node "node1"
  {
    fastname "$host"
    pools ""
    resource disk "/mydisk1/tmp" {pools ""}
    resource scratchdisk "/myscratchdisk1/tmp" {pools ""}
    instances 30
  }
  node "node31"
  {
    fastname "machineA.domain.com"
    pools ""
    resource disk "/mydisk2/tmp" {pools ""}
    resource scratchdisk "/myscratchdisk2/tmp" {pools ""}
    instances 10
  }
  node "node41"
  {
    fastname "machineB.domain.com"
    pools ""
    resource disk "/mydisk3/tmp" {pools ""}
    resource scratchdisk "/myscratchdisk3/tmp" {pools ""}
  }
}

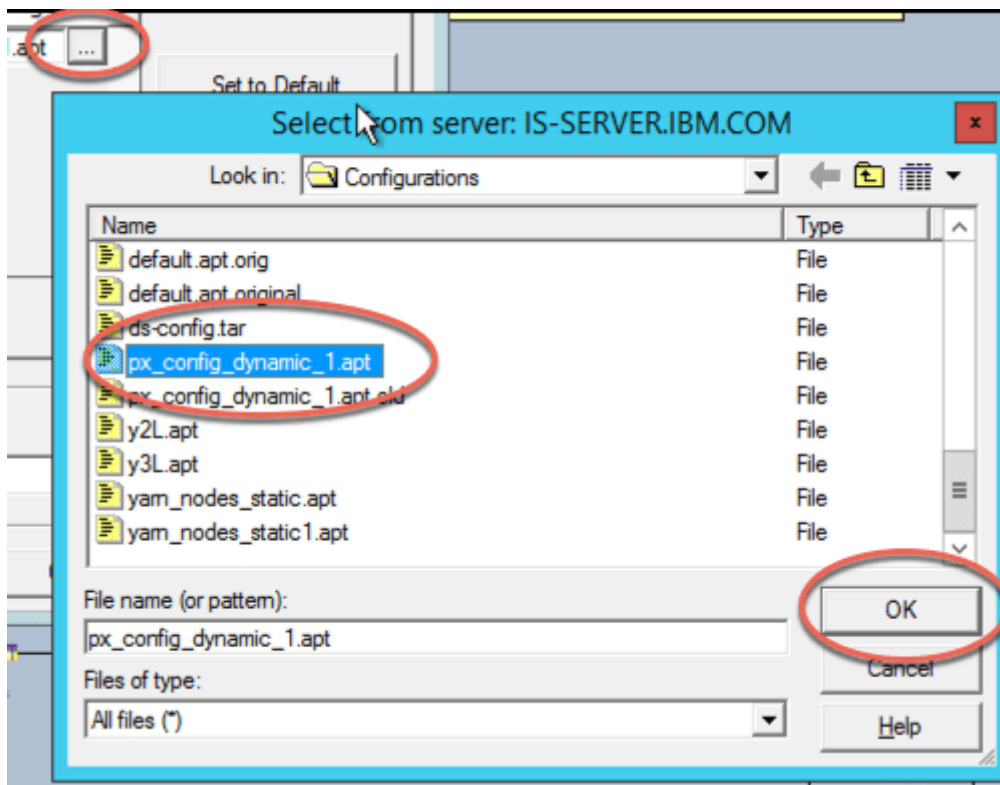
```

5. In the Job Run Options window, use a dynamic configuration file for your processing. Click the field that defines the configuration file to show the **Options** icon.

A file-browser window is displayed and the `px_config_dynamic_1.appt` file is selected.

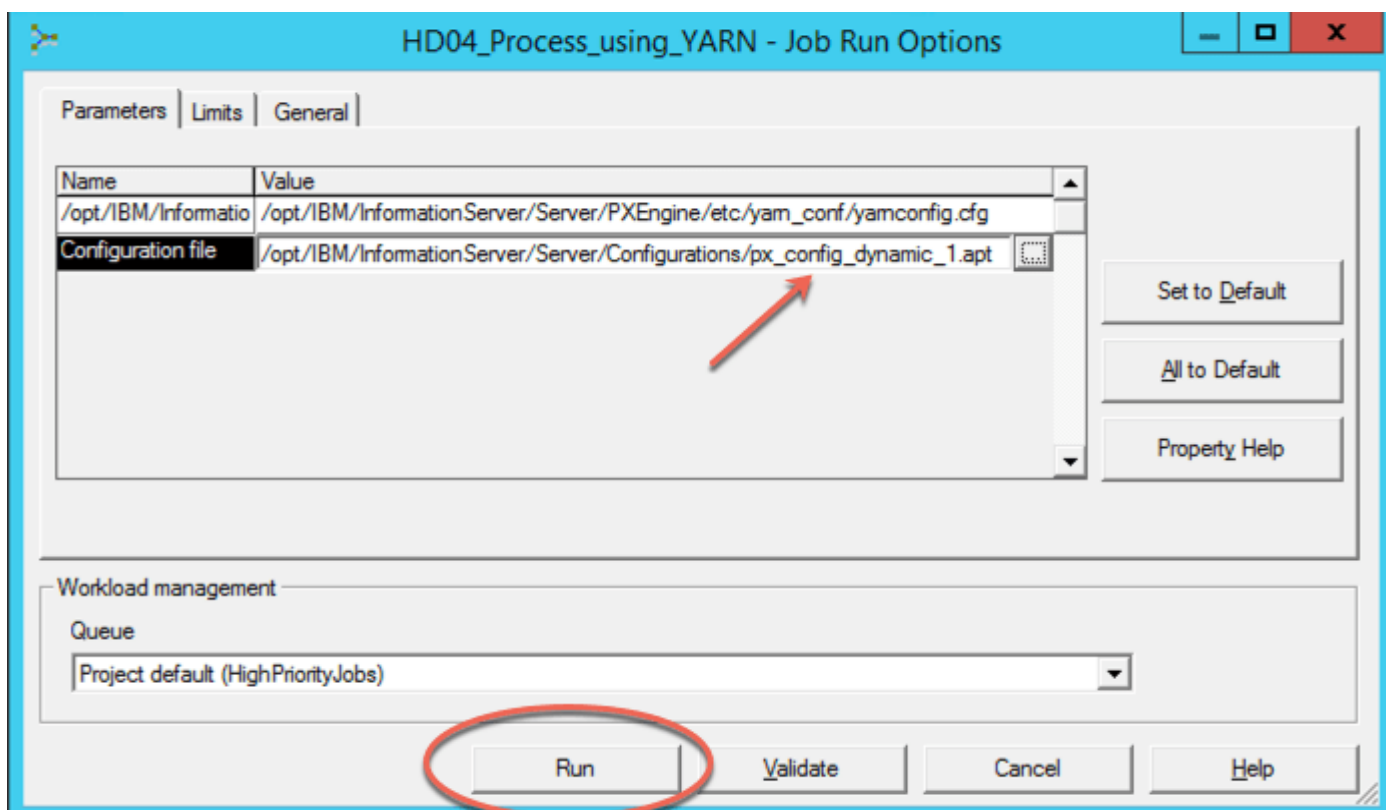


6. Click **OK**.

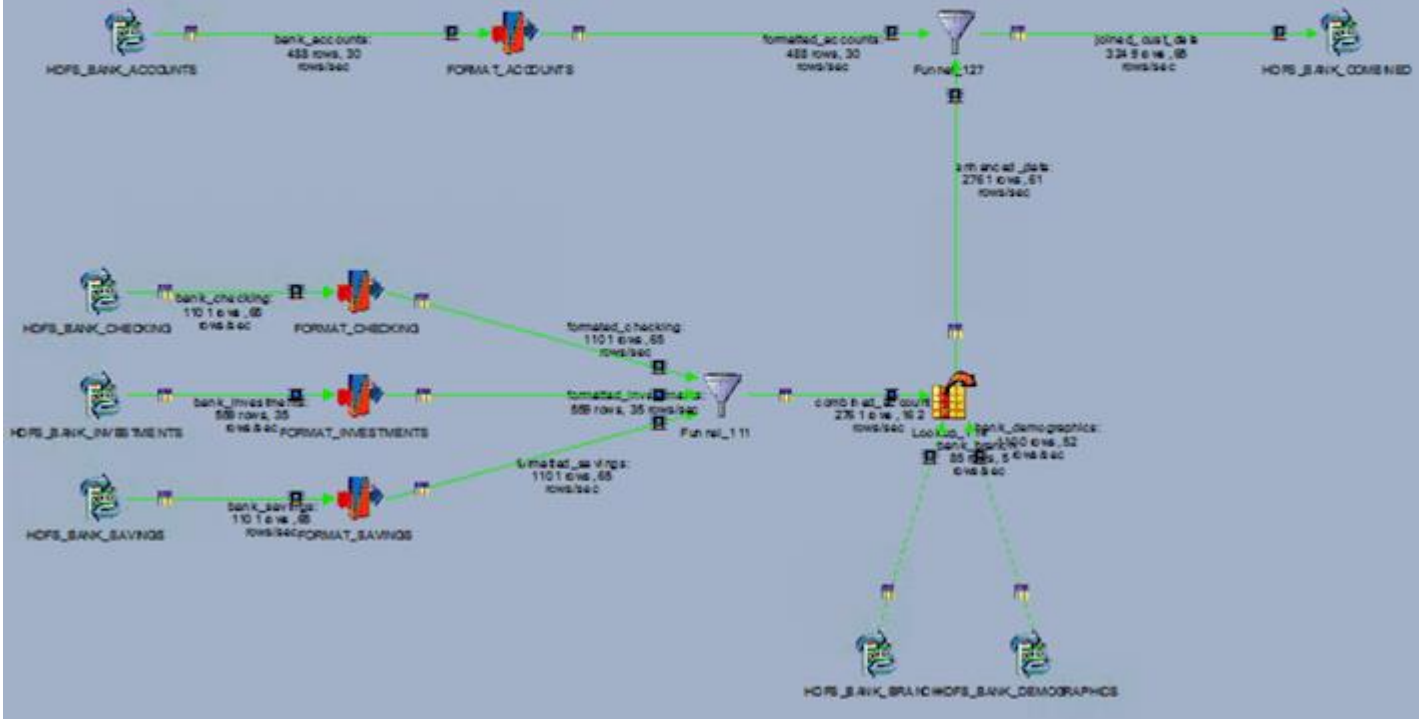


7. Stellen Sie sicher, dass Sie die richtige Konfigurationsdatei verwenden, und klicken Sie auf **Ausführen**.

Wenn der Auftrag abgeschlossen ist, werden alle Auftragsverknüpfungen grün und zeigen die Anzahl der Zeilen auf jedem Link an.



Push Processing onto HDFS and run using YARN



8. Klicken Sie auf die Mitte des unteren Teils der Seite, um das Auftragsprotokollfenster zu erhöhen.

9. Achten Sie auf diese Einträge:

```
Starting Job HD04_Process_using_YARN. (...)
Environment variable settings: (...)
Parallel job initiated
OSH script (...)
Parallel job default NLS map UTF-8, default locale OFF
main_program: IBM InfoSphere DataStage Enterprise Edition 11.5.0.8169 (...)
main_program: The open files limit is 1024; raising to 4096.
main_program: conductor uname: -s=Linux; -r=2.6.32-642.13.1.el6.x86_64; -v=#1 SMP Wed Nov 23 16:03:01 EST 2016; -n=1s-serve...
main_program: orchgeneral: loaded (...)
main_program: Parallel Engine running in YARN execution mode.
main_program: IPv6 isn't currently supported by Hadoop/YARN and the required environment variable APT_USE_IPV4 is not set. It ...
HDFS_BANK_ACCOUNTS: Accessing file via WebHDFS file system.
HDFS_BANK_CHECKING: Accessing file via WebHDFS file system.
HDFS_BANK_INVESTMENTS: Accessing file via WebHDFS file system.
HDFS_BANK_SAVINGS: Accessing file via WebHDFS file system.
HDFS_BANK_DEMOGRAPHICS: Accessing file via WebHDFS file system.
HDFS_BANK_BRANCH: Accessing file via WebHDFS file system.
HDFS_BANK_COMBINED: The connector was configured to run in parallel on 3 nodes, but the Read/Write mode is Read/Write sin...
HDFS_BANK_COMBINED: Accessing file via WebHDFS file system
main_program: APT configuration file: /opt/IBM/InformationServer/Server/Configurations/px_config_dynamic_1.apt (...)
```

Zusammenfassung

Sie haben einen herkömmlichen Data Warehouse ETL-Auftrag ausgeführt und Daten von Data Warehouse nach Hadoop verschoben. Anschließend haben Sie eine Data Warehouse ETL mit Hadoop-Daten ausgeführt und die ETL-Verarbeitung in Hadoop mithilfe von YARN ausgeführt.