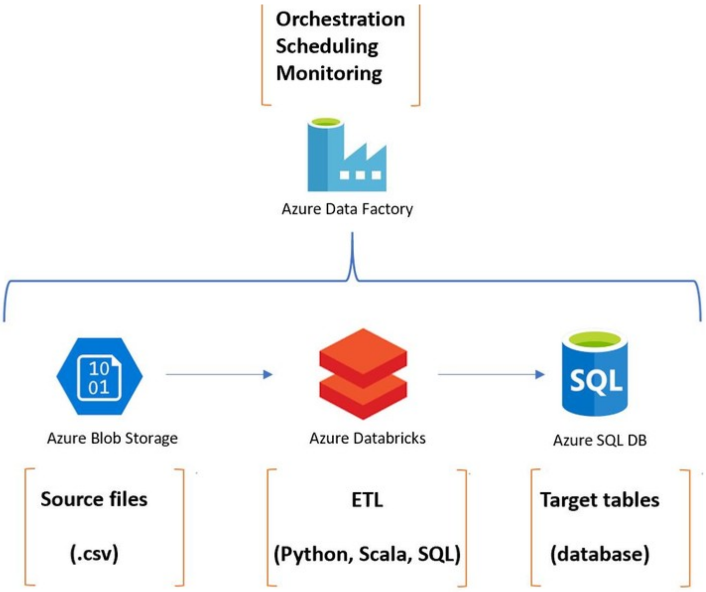


Azure Databricks with Azure Data Factory

ETL Modernization using Azure Databricks and Azure Data Factory

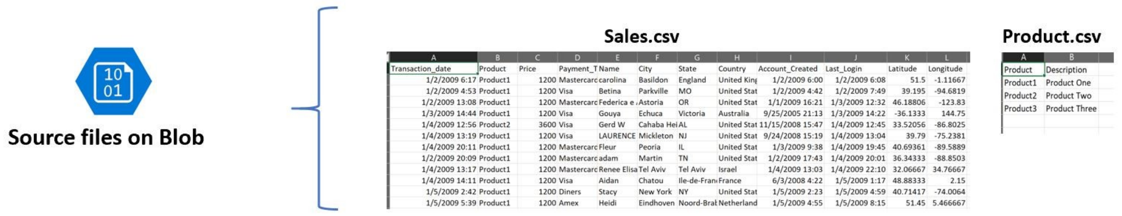
Published on February 26, 2019

Rahul Athale
Director, Azure Analytics & AI, Customer Success - Worldwide Commercial Business

6 articles + Follow

A simple example depicting commonly used ETL activities like: (1) Joining Two Datasets (2) Renaming Column (3) Creating a new column with a computed value.

Source Data:



Sales.csv

A	B	C	D	E	F	G	H	I	J	K	L
Transaction_date	Product	Price	Payment_T	Name	City	State	Country	Account_Created	Last_Login	Latitude	Longitude
1/2/2009 6:17	Product1	1200	Mastercarc	carolina	Basildon	England	United King	1/2/2009 6:00	1/2/2009 6:08	51.5	-1.11667
1/2/2009 4:53	Product1	1200	Visa	Betina	Parkville	MO	United Stat	1/2/2009 4:42	1/2/2009 7:49	39.195	-94.6819
1/2/2009 13:08	Product1	1200	Mastercarc	Federica e /	Astoria	OR	United Stat	1/1/2009 16:21	1/3/2009 12:32	46.18806	-123.83
1/3/2009 14:44	Product1	1200	Visa	Gouya	Echuca	Victoria	Australia	9/25/2005 21:13	1/3/2009 14:22	-36.1333	144.75
1/4/2009 12:56	Product2	3600	Visa	Gerd W	Cahaba Hei	AL	United Stat	11/15/2008 15:47	1/4/2009 12:45	33.52056	-86.8025
1/4/2009 13:19	Product1	1200	Visa	LAURENCE	Mickleton	NJ	United Stat	9/24/2008 15:19	1/4/2009 13:04	39.79	-75.2381
1/4/2009 20:11	Product1	1200	Mastercarc	Fleur	Peoria	IL	United Stat	1/3/2009 9:38	1/4/2009 19:45	40.69361	-89.5889
1/2/2009 20:09	Product1	1200	Mastercarc	adam	Martin	TN	United Stat	1/2/2009 17:43	1/4/2009 20:01	36.34333	-88.8503
1/4/2009 13:17	Product1	1200	Mastercarc	Renee Elisa	Tel Aviv	Tel Aviv	Israel	1/4/2009 13:03	1/4/2009 22:10	32.06667	34.76667
1/4/2009 14:11	Product1	1200	Visa	Aidan	Chatou	Ile-de-Fran	France	6/3/2008 4:22	1/5/2009 1:17	48.88333	2.15
1/5/2009 2:42	Product1	1200	Diners	Stacy	New York	NY	United Stat	1/5/2009 2:23	1/5/2009 4:59	40.71417	-74.0064
1/5/2009 5:39	Product1	1200	Amex	Heidi	Eindhoven	Noord-Bral	Netherland	1/5/2009 4:55	1/5/2009 8:15	51.45	5.466667

Product.csv

A	B
Product	Description
Product1	Product One
Product2	Product Two
Product3	Product Three

Two .csv files on Azure Blob - "Sales.csv" & "Product.csv"

ETL Activities:

Use an Azure Databricks Python Notebook to perform these transformation tasks.

- *Mount an Azure Blob Storage container as DBFS*

```
dbutils.fs.mount(
    source = "wasbs://"+getArgument("filepath")+"@"+storageName+".blob.co
    mount_point = "/mnt/adfdata",
    extra_configs = {"fs.azure.account.key."+storageName+".blob.core.wind
                     accessKey})
```

- *Access files in your container as if they were local files*

```
%python

df_sales = sqlContext.read.format('csv').options(header='true', inferSche
df_products = sqlContext.read.format('csv').options(header='true', inferS
```

- *Join two dataframes ; Rename column and sort ; Create ne*

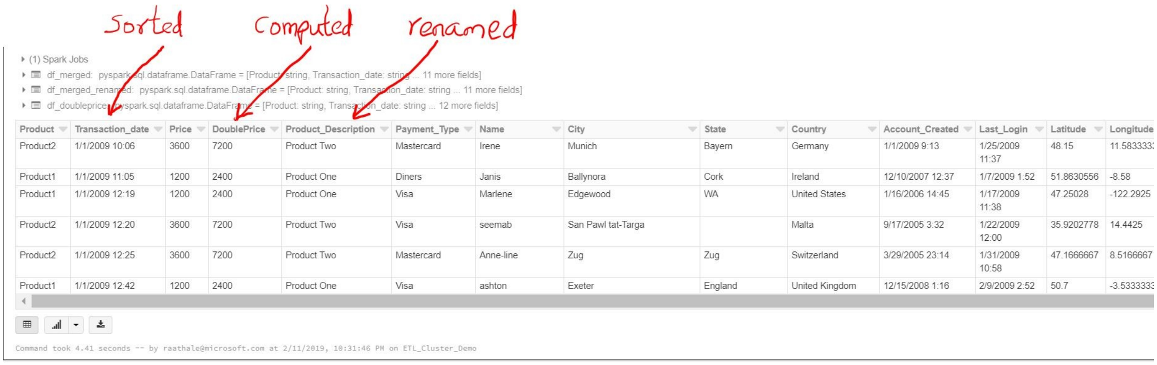
```
from pyspark.sql.functions import col


# Join two dataframes on a column called "Product"
df_merged = df_sales.join(df_products, ["Product"])


# Rename column to new name, sort on one column
df_merged_renamed = (df_merged
                    .withColumnRenamed("Description", "Product_Description")
                    .sort("Transaction_date")
                    )


# create new column with Computed value and call it doublePrice
df_doubleprice = (df_merged_renamed
                 .withColumn("DoublePrice", col("Price") * 2)
                 .select("Product", "Transaction_date", "Price", "DoublePrice", "Product_Descriptic


display(df_doubleprice)
```



```
(1) Spark Jobs
df_merged:  pyspark.sql.dataframe.DataFrame = [Product: string, Transaction_date: string ... 11 more fields]
df_merged_renamed:  pyspark.sql.dataframe.DataFrame = [Product: string, Transaction_date: string ... 11 more fields]
df_doubleprice:  pyspark.sql.dataframe.DataFrame = [Product: string, Transaction_date: string ... 12 more fields]
```

Product	Transaction_date	Price	DoublePrice	Product_Description	Payment_Type	Name	City	State	Country	Account_Created	Last_Login	Latitude	Longitude
Product2	1/1/2009 10:06	3600	7200	Product Two	Mastercard	Irene	Munich	Bayern	Germany	1/1/2009 9:13	1/25/2009 11:37	48.15	11.583333
Product1	1/1/2009 11:05	1200	2400	Product One	Diners	Janis	Ballynora	Cork	Ireland	12/10/2007 12:37	1/7/2009 1:52	51.8630556	-8.58
Product1	1/1/2009 12:19	1200	2400	Product One	Visa	Marlene	Edgewood	WA	United States	1/16/2006 14:45	1/17/2009 11:38	47.25028	-122.2925
Product2	1/1/2009 12:20	3600	7200	Product Two	Visa	seemab	San Pawl tat-Targa		Malta	9/17/2005 3:32	1/22/2009 12:00	35.9202778	14.4425
Product2	1/1/2009 12:25	3600	7200	Product Two	Mastercard	Anne-line	Zug	Zug	Switzerland	3/29/2005 23:14	1/31/2009 10:58	47.1666667	8.5166667
Product1	1/1/2009 12:42	1200	2400	Product One	Visa	ashton	Exeter	England	United Kingdom	12/15/2008 1:16	2/9/2009 2:52	50.7	-3.533333

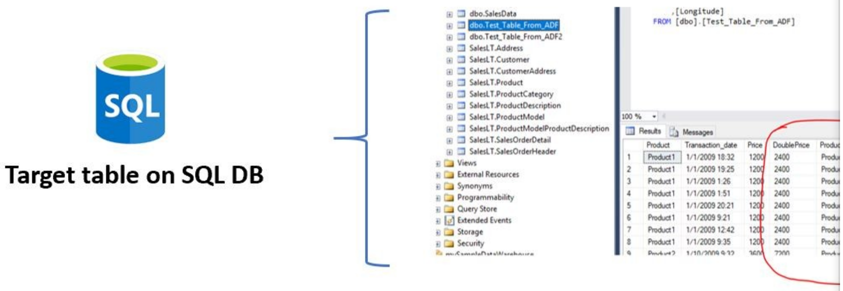
```
Command took 4.41 seconds -- by raathale@microsoft.com at 2/11/2019, 10:31:46 PM on ETL_Cluster_Demo
```

Target Table:

```
jdbcUrl = "jdbc:sqlserver://{0}:{1};database={2};encrypt=true;trustServe
```

- *Load/Write ETL output to Table in SQL Server*

```
db_target_table = getArgument("target_tablename")
df_doubleprice.write.jdbc(jdbcUrl, db_target_table, properties=connectio
```



```
      ,[Longitude]
  FROM [dbo].[Test_Table_From_ADF]
```

	Product	Transaction_date	Price	DoublePrice	Produ
1	Product1	1/1/2009 18:32	1200	2400	Produ
2	Product1	1/1/2009 19:25	1200	2400	Produ
3	Product1	1/1/2009 1:26	1200	2400	Produ
4	Product1	1/1/2009 1:51	1200	2400	Produ
5	Product1	1/1/2009 20:21	1200	2400	Produ
6	Product1	1/1/2009 9:21	1200	2400	Produ
7	Product1	1/1/2009 12:42	1200	2400	Produ
8	Product1	1/1/2009 9:35	1200	2400	Produ
9	Product2	1/10/2009 9:32	3600	7200	Produ

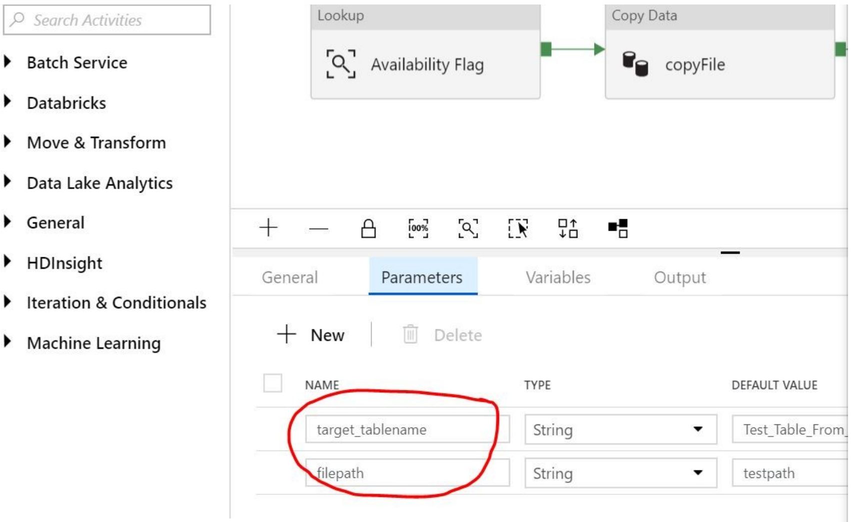
A database table *(with name based on value of parameter "targe* created or updated in the SQL Server database. This table is the result set of joining "Sales.csv" & "Product.csv" + renaming of "Description" column + creation of new computed column "DoublePrice".

Orchestration using Azure Data Factory:

Pipeline:

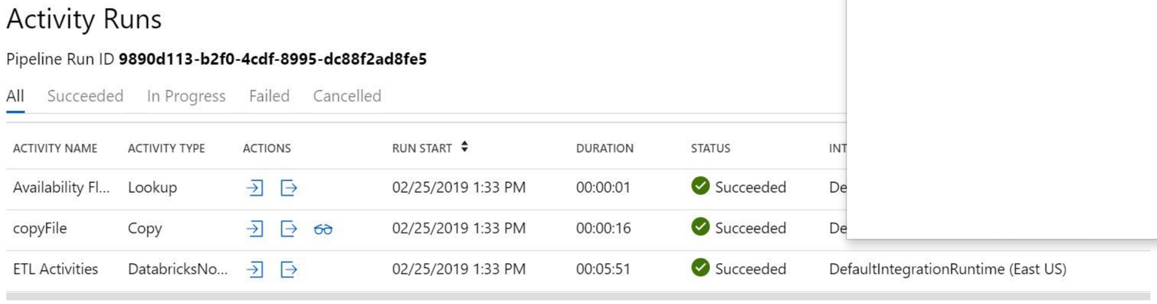
- Lookup activity: to check for source data availability at directory location
- Copy activity: to copy processed files into different directory location
- Databricks Notebook activity: ETL tasks and load data into SQL Database

Two parameters are passed from Azure Data Factory to Azure Databricks - "filepath" & "target_tablename"



NAME	TYPE	DEFAULT VALUE
target_tablename	String	Test_Table_From_
filepath	String	testpath

Activity Runs:

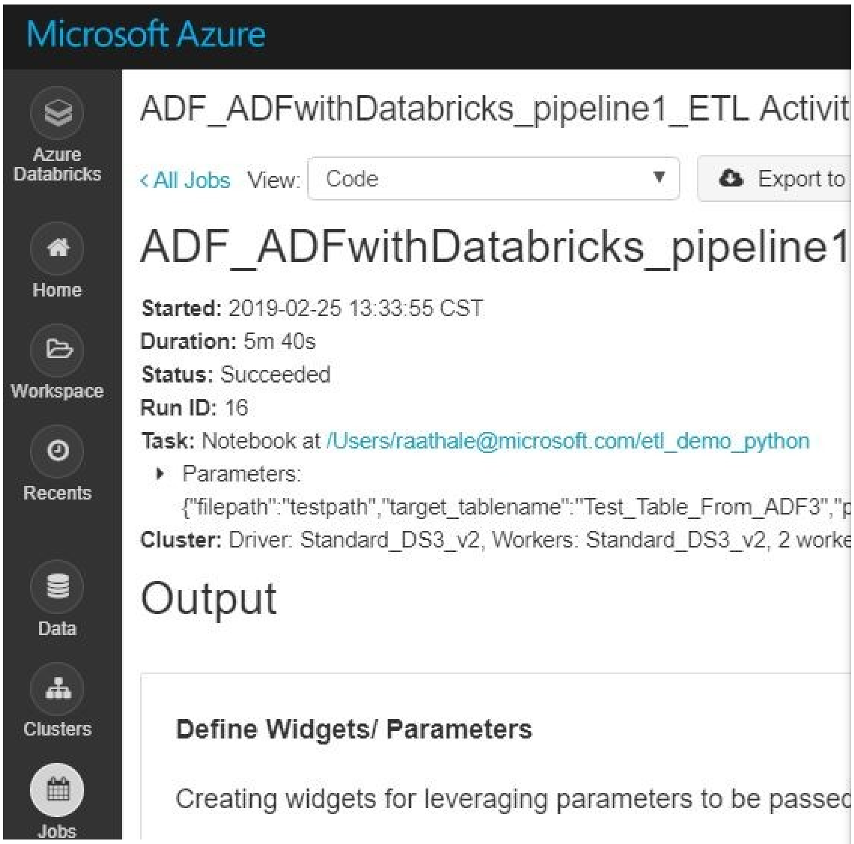


Activity Runs

Pipeline Run ID **9890d113-b2f0-4cdf-8995-dc88f2ad8fe5**

ACTIVITY NAME	ACTIVITY TYPE	ACTIONS	RUN START	DURATION	STATUS	INT
Availability Fl...	Lookup		02/25/2019 1:33 PM	00:00:01	✓ Succeeded	De
copyFile	Copy		02/25/2019 1:33 PM	00:00:16	✓ Succeeded	De
ETL Activities	DatabricksNo...		02/25/2019 1:33 PM	00:05:51	✓ Succeeded	DefaultIntegrationRuntime (East US)

Databricks Job:



ADF_ADFwithDatabricks_pipeline1_ETL Activit

ADF_ADFwithDatabricks_pipeline1

Started: 2019-02-25 13:33:55 CST
Duration: 5m 40s
Status: Succeeded
Run ID: 16
Task: Notebook at /Users/raathale@microsoft.com/etl_demo_python
Parameters:
{"filepath":"testpath","target_tablename":"Test_Table_From_ADF3","p
Cluster: Driver: Standard_DS3_v2, Workers: Standard_DS3_v2, 2 worke

Output

Define Widgets/ Parameters

Creating widgets for leveraging parameters to be passed

#Azure #AzureDatabricks #AzureDataFactory #Microsoft

Report this

Published by

Rahul Athale
Director, Azure Analytics & AI, Customer Success - Worldwide Commercial Business
Published • 1yr

6 articles + Follow

#Azure #AzureDatabricks #AzureDataFactory #Microsoft

Reactions

...

7 Comments

Most Relevant

Add a comment...

Hayk 'Hike' C. • 3rd+
Let's Talk Data | *Hiring*
1y ...

How do you get all of the data into azure blob storage?

· 1 Like | · 1 Reply

you can upload into Azure Blob using multiple options. That isn't the us

Yago Luiz • 3rd+
Senior Backend Developer at Wiz Soluções | Master's degree at University of Brasília

Sidney Oliveira Raphael Brito

· 3 Likes

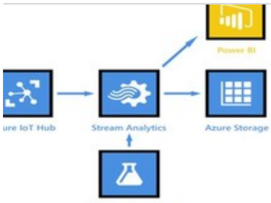
Load more comments

Rahul Athale

Director, Azure Analytics & AI, Customer Success - Worldwide Commerc

+ Follow

n Rahul Athale



Stream Analytics & IoT

e on LinkedIn

Azure Machine Learning

	Using AML Studio	Using Linux DSVM
Prepare data	✓	✓
Build a model	✓	✓
Operationalize a model	✓	✓
Consume a model	✓	✓

Azure Machine Learning – a simple example

Rahul Athale on LinkedIn

Kafka + Spark Streaming + ADLS

Rahul Athale on LinkedIn

Azure SQL DW - cost saving
scalable DWUs

Rahul Athale on LinkedIn

icles

Messaging