


1. Deployment of downloading, ETL and Integratiton process

System information

Spark Cluster	HDInsight
Cluster OS	Ubuntu
Cloud Storage	Azure blob

1. Create Azure blob and container

Azure Blob

	Storage account: east8aota8blob Container: aota														
	<div><div>Home > east8aota8blob ></div><div><div> aota ... Container</div><div><input type="text" value="Search (Ctrl+/)"/></div><div><div>Overview</div><div>Diagnose and solve problems</div><div>Access Control (IAM)</div><div>Settings</div><div>Shared access tokens</div><div>Access policy</div><div>Properties</div><div>Metadata</div></div><div><div>Upload</div><div>Change access level</div><div><input type="text" value="Search blobs by prefix (case-sensitive)"/></div><div>Add filter</div><table><thead><tr><th>Name</th><th>Modified</th></tr></thead><tbody><tr><td><input type="checkbox"/> dwsource</td><td></td></tr><tr><td><input type="checkbox"/> <u>integrated</u></td><td></td></tr><tr><td><input type="checkbox"/> logs</td><td></td></tr><tr><td><input type="checkbox"/> <u>processed</u></td><td></td></tr><tr><td><input type="checkbox"/> <u>raw</u></td><td></td></tr><tr><td><input type="checkbox"/> .</td><td></td></tr></tbody></table></div></div></div>	Name	Modified	<input type="checkbox"/> dwsource		<input type="checkbox"/> <u>integrated</u>		<input type="checkbox"/> logs		<input type="checkbox"/> <u>processed</u>		<input type="checkbox"/> <u>raw</u>		<input type="checkbox"/> .	
Name	Modified														
<input type="checkbox"/> dwsource															
<input type="checkbox"/> <u>integrated</u>															
<input type="checkbox"/> logs															
<input type="checkbox"/> <u>processed</u>															
<input type="checkbox"/> <u>raw</u>															
<input type="checkbox"/> .															

2. Create HDinsight cluster

Here are some details about setting up the HDinsight cluster

Cluster Configuration	<div><div>+ Add application</div><table><tr><th>Node type</th><th>Node size</th><th>Number of ...</th><th>Estimated cost/h...</th></tr><tr><td>Head node</td><td>D12 v2 (4 Cores, 28 GB RAM), 0.37 USD/hour</td><td>2</td><td>0.75 USD</td></tr><tr><td>Zookeeper node</td><td>A2 v2 (2 Cores, 4 GB RAM), 0.13 USD/hour</td><td>3</td><td>0.00 (FREE)</td></tr><tr><td>Worker node</td><td>D12 v2 (4 Cores, 28 GB RAM), 0.37 USD/hour</td><td>2</td><td>0.75 USD</td></tr></table><div><input type="checkbox"/> Enable autoscale Learn More</div></div>	Node type	Node size	Number of ...	Estimated cost/h...	Head node	D12 v2 (4 Cores, 28 GB RAM), 0.37 USD/hour	2	0.75 USD	Zookeeper node	A2 v2 (2 Cores, 4 GB RAM), 0.13 USD/hour	3	0.00 (FREE)	Worker node	D12 v2 (4 Cores, 28 GB RAM), 0.37 USD/hour	2	0.75 USD
Node type	Node size	Number of ...	Estimated cost/h...														
Head node	D12 v2 (4 Cores, 28 GB RAM), 0.37 USD/hour	2	0.75 USD														
Zookeeper node	A2 v2 (2 Cores, 4 GB RAM), 0.13 USD/hour	3	0.00 (FREE)														
Worker node	D12 v2 (4 Cores, 28 GB RAM), 0.37 USD/hour	2	0.75 USD														
Add additional storage	<p>When creating the cluster, we add an Azure storage account, 'east8aota8blob', to be the additional storage. In this way, our Spark program can freely access it through some settings.</p> <div><div>Additional Azure Storage</div><div>Link additional Azure Storage accounts to the cluster.</div><div>Account name</div><div>east8aota8blob</div><div>Add Azure Storage</div></div>																
blobfuse	<p>On Cluster, we installed blobfuse (on linux) which allows to mount Blob storage ('east8aota8blob'), as a local file system.</p> <p>In such a way, our python script can download the source data and directly save to it.</p>																

3. Deploy code HDinsight

- 1) SSH to HDinsight Cluster
- 2) Commands

----- ssh-----

```
ssh sshuser@cluster-0531-ssh.azurehdinsight.net
```

```
----- install moudle -=====
```

```
sudo pip3 install pyspark
```

```
----- set up folder for aota
```

```
-----  
sudo mkdir /app  
sudo chown sshuser /app  
mkdir /app/aota
```

```
sudo mkdir /data  
sudo chown sshuser /data
```

```
#create folder:  
sudo mkdir /data/aota  
sudo mkdir /data/aota/raw  
sudo mkdir /data/aota/logs  
sudo mkdir /data/aota/processed  
sudo mkdir /data/aota/raw
```

```
sudo mkdir /data/aota/integrated
```

```
#change owner  
sudo chown sshuser /data/aota/logs --recursive
```

```
sudo chown sshuser /data --recursive
```

```
sudo chown sshuser /app --recursive
```

```
----- how to upload code-----
```

```
upload to hdfs:/tmp of primary storage
```

```
----- download code -----
```

```
hdfs dfs -ls /  
hdfs dfs -get /tmp /app/aota
```

```
----- after copy logs to hdfs -----  
hdfs dfs -put /data/aota/logs /tmp/logs
```

```
----- export -----  
export SPARK_HOME=/usr/lib/spark  
export sshuser_HOME=/usr/lib/sshuser
```

```
#sudo nano $SPARK_HOME/conf/spark-defaults.conf

--- add user to root group ??? -----

sudo usermod -G root sshuser
```

4. Running

Process	Timing
Download	Start: 2022-05-29_17:43:29 End: 2022-05-29_17:43:59 Duration: 30 seconds
ETL (cleanse/Transform)	Start: 2022-05-29_17:44:40 End: 2022-05-29_18:14:42 Duration: 30 minutes
Integration process	Start: 2022-05-31_19:44:00 End: 2022-05-31_19:55:46 Duration: 12 minutes

2. Deployment of DW-ETL

Preparation:

1. Convert 'integration layer' parquet files to CSV files
2. Upload CSV files to a S3 bucket using AWS cli command.

Timing: it took 30 minutes to load csv files to S3 bucket

Timing	Start: 22:09:48.40 End: 21:39:21.61
--------	--

	Duration: 30 minutes
--	----------------------

- Load data to Snowflake,
- 1) Create database and tables in Snowflakes
 - 2) Create stage
 - 3) Using 'copy' command to load data from S3 bucket to Snowflake tables

Row count loaded	59 millions																																																																																								
Time	1 minute																																																																																								
Screen shot	<div><div><div><div>▶ Run</div><div><input type="checkbox"/> All Queries</div><div>Saved 2 minutes ago</div></div><div><div>ACCOUNTADMIN</div><div>COMPUTE_WH (XS)</div><div>AOT</div><div>PUBLIC</div><div>...</div></div></div><div><pre>110 arr_time string , 111 scheduled_arrtime string , 112 arr_delay integer , 113 year integer 114); 115 116 copy into fact_flight from @STG_SOURCE_FACT_FLIGHT file_format =csv; 117 118 copy into fact_flight from @stg_source_fact_flight file_format =csv; 119</pre></div><div><div>Results</div><div>Data Preview</div><div>Open History</div></div><div><div>✓ Query ID</div><div>SQL</div><div>1m</div><div>74 rows</div><div>Filter result...</div><div>Download</div><div>Copy</div><div>Columns</div></div><table><tr><th>Row</th><th>file</th><th>status</th><th>rows_parsed</th><th>rows_loaded</th><th>error_limit</th><th>errors_seen</th><th>first_error</th><th>first_error_line</th><th>first_error_chara</th><th>first_error_colur</th></tr><tr><td>67</td><td>s3://capston...</td><td>LOADED</td><td>874569</td><td>874569</td><td>1</td><td>0</td><td>NULL</td><td>NULL</td><td>NULL</td><td>NULL</td></tr><tr><td>68</td><td>s3://capston...</td><td>LOADED</td><td>872525</td><td>872525</td><td>1</td><td>0</td><td>NULL</td><td>NULL</td><td>NULL</td><td>NULL</td></tr><tr><td>69</td><td>s3://capston...</td><td>LOADED</td><td>873017</td><td>873017</td><td>1</td><td>0</td><td>NULL</td><td>NULL</td><td>NULL</td><td>NULL</td></tr><tr><td>70</td><td>s3://capston...</td><td>LOADED</td><td>662539</td><td>662539</td><td>1</td><td>0</td><td>NULL</td><td>NULL</td><td>NULL</td><td>NULL</td></tr><tr><td>71</td><td>s3://capston...</td><td>LOADED</td><td>685671</td><td>685671</td><td>1</td><td>0</td><td>NULL</td><td>NULL</td><td>NULL</td><td>NULL</td></tr><tr><td>72</td><td>s3://capston...</td><td>LOADED</td><td>649506</td><td>649506</td><td>1</td><td>0</td><td>NULL</td><td>NULL</td><td>NULL</td><td>NULL</td></tr><tr><td>73</td><td>s3://capston...</td><td>LOADED</td><td>649506</td><td>649506</td><td>1</td><td>0</td><td>NULL</td><td>NULL</td><td>NULL</td><td>NULL</td></tr></table></div>	Row	file	status	rows_parsed	rows_loaded	error_limit	errors_seen	first_error	first_error_line	first_error_chara	first_error_colur	67	s3://capston...	LOADED	874569	874569	1	0	NULL	NULL	NULL	NULL	68	s3://capston...	LOADED	872525	872525	1	0	NULL	NULL	NULL	NULL	69	s3://capston...	LOADED	873017	873017	1	0	NULL	NULL	NULL	NULL	70	s3://capston...	LOADED	662539	662539	1	0	NULL	NULL	NULL	NULL	71	s3://capston...	LOADED	685671	685671	1	0	NULL	NULL	NULL	NULL	72	s3://capston...	LOADED	649506	649506	1	0	NULL	NULL	NULL	NULL	73	s3://capston...	LOADED	649506	649506	1	0	NULL	NULL	NULL	NULL
Row	file	status	rows_parsed	rows_loaded	error_limit	errors_seen	first_error	first_error_line	first_error_chara	first_error_colur																																																																															
67	s3://capston...	LOADED	874569	874569	1	0	NULL	NULL	NULL	NULL																																																																															
68	s3://capston...	LOADED	872525	872525	1	0	NULL	NULL	NULL	NULL																																																																															
69	s3://capston...	LOADED	873017	873017	1	0	NULL	NULL	NULL	NULL																																																																															
70	s3://capston...	LOADED	662539	662539	1	0	NULL	NULL	NULL	NULL																																																																															
71	s3://capston...	LOADED	685671	685671	1	0	NULL	NULL	NULL	NULL																																																																															
72	s3://capston...	LOADED	649506	649506	1	0	NULL	NULL	NULL	NULL																																																																															
73	s3://capston...	LOADED	649506	649506	1	0	NULL	NULL	NULL	NULL																																																																															