

# MSBA Financial Group AWS Cloud Integration

Younjoo

Lee



[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

# Data Flow Diagram

Outlining the architecture of the project

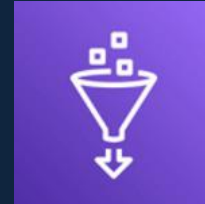
## 3 Disparate Data Sources:

- 1) Datacorp Vendor: datacorp\_financials for balance sheets and income statements (.csv file)
- 2) Analyst A: msba\_fg\_ratios for financial ratios (.csv file)
- 3) Analyst B: msba\_fg\_bankruptcy for bankruptcy data (.txt file)

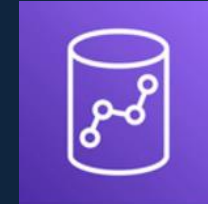
## 1) Data Lake AWS S3



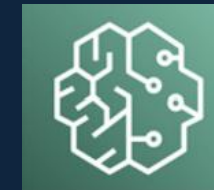
## 1.5) ETL AWS Glue



## 2) Data Warehouse AWS Redshift



## 3) Machine Learning AWS Sagemaker



Companies to predict bankruptcy (.csv file)

Results: Bankruptcy  
Prediction Data (.csv file)

# 1) Data Lake | AWS S3

aws

Services

Search [Alt+S]

Global

voclabs/user2745682=Younjoo\_Lee @ 5050-7925-0130

Amazon S3

Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

AWS Organizations settings

Feature spotlight 7

AWS Marketplace for S3

Amazon S3

Buckets

msba-fg-101

msba-fg-101

Info

Objects

Properties

Permissions

Metrics

Management

Access Points

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Find objects by prefix

< 1 >

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	data_files/	Folder	-	-	-
<input type="checkbox"/>	prediction/	Folder	-	-	-

3

# 2) Data Warehouse | AWS Redshift

The screenshot displays the AWS Redshift Query Editor interface. At the top, the AWS logo and navigation bar are visible, including the 'Services' menu and a search bar containing 'glue'. The main header shows the 'Amazon Redshift' logo and the 'Query editor' title. Below this, there are tabs for 'Editor', 'Query history', 'Saved queries', and 'Scheduled queries'. The 'Editor' tab is active, showing a SQL query editor with a status bar at the top indicating 'Status: Connected', 'database: dev', 'user: msbauser', and a 'Change connection' button. The query editor has a toolbar with icons for undo, redo, and other editing functions. The SQL query being edited is:

```
1 CREATE TABLE financials_combined
2 (
3   Company_ID          VARCHAR(10) NOT NULL
4 )
```

On the left side, there is a 'Resources' panel with a search bar and a list of databases and schemas. The 'dev' database is selected, and the 'public' schema is chosen. Below this, a list of tables is shown, including 'bankruptcy' and 'financials\_combined'. The 'financials\_combined' table is expanded, showing its columns: 'company\_id', 'liability\_to\_equity', 'net\_income\_to\_total\_assets', 'working\_capital\_to\_total\_assets', 'net\_profit\_before\_tax\_to\_paid\_in\_capital', 'operating\_profit\_per\_share', and 'net\_worth\_to\_assets'. At the bottom of the editor, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear', along with a 'Send feedback' link. The bottom status bar shows 'Query results' and 'Table details' tabs, with 'Table details' currently selected.

# 1.5) ETL | AWS Glue

Services  [Alt+S] N. Virginia ▼ voclabs/user2745682=Younjoo\_Lee @ 5050-7925-0130 ▼

**AWS Glue** ✕

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

► Data Catalog

► Data Integration and ETL

► Legacy pages

What's New

Documentation

AWS Marketplace

☒ Enable compact mode

☒ Enable new navigation

**Final Project Job FG**

Last modified on 10/5/2023, 10:24:58 AM ☒ Try new UI

Actions ▼

Save

Run

Visual | Script | Job details | Runs | Data quality New | Schedules | Version Control

+

Data source - S3 bucket  
S3 Financials

Data source - S3 bucket  
S3 Ratios

Transform - Join  
Join

Transform - Change Sche...  
Change Schema

Data target - Amazon Re...  
Amazon Redshift

**No node selected**  
Choose a node from the graph to view its configuration properties.

5

# 1.5) ETL | AWS Glue - Automate

aws

Services

Search

[Alt+S]

N. Virginia

voclabs/user2745682=Younjoo\_Lee @ 5050-7925-0130

AWS Glue

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

► Data Catalog

► Data Integration and ETL

► Legacy pages

What's New

Documentation

AWS Marketplace

Enable compact mode

Enable new navigation

Final Project Job FG

Last modified on 10/5/2023, 10:24:58 AM

Try new UI

Actions

Save

Run

Visual

Script

Job details

Runs

Data quality New

Schedules

Version Control

Schedules

Filter schedules

Actions

Create schedule

< 1 >

No schedules

No schedules available

Create schedule

6

# 3) Machine Learning | AWS Sagemaker - Canvas

aws

Services

Search

[Alt+S]

N. Virginia

voclabs/user2745682=Younjoo\_Lee @ 5050-7925-0130

SageMaker dashboard

Search

▼ JumpStart

Foundation models NEW

Computer vision models

Natural language processing models

► Governance

► Ground Truth

▼ Notebook

Notebook instances

Git repositories

► Processing

► Training

► Inference

► Edge Manager

► Augmented AI

Sagemaker geospatial capability is now generally available in us-west-2

Amazon SageMaker geospatial capabilities make it easier for data scientists and machine learning (ML) engineers to build, train, and deploy ML models faster using geospatial data.

Learn more

Amazon SageMaker > Domains > Domain: msba-sagemaker

msba-sagemaker

Domain details

Configure and manage the domain.

User profiles | Space management | Environment | Domain settings

User profiles Info

A user profile represents a single user within a domain. It is the main way to reference a user for the purposes of sharing, reporting, and other use

Search users

Name	Modified on	Created on
msba-data-analyst	Sep 28, 2023 05:12 UTC	Sep 28, 2023 05:12 UTC

Personal apps

Studio

Canvas

TensorBoard

Profiler

Collaborative

Spaces

Launch

7

# 3a) Creating a Prediction Model

Amazon SageMaker  
Canvas

Ready-to-use models

My models

Shared models

Datasets

Automations

Help

Log out

My models / bankruptcy\_prediction\_model / Version 1

+ Add version

↺

⋮

SelectBuildAnalyzePredict

Model status

Accuracy ⓘ  
**97.067%**

F1 ⓘ Optimization metric  
0.535

Predict

The model predicts the correct Bankrupt 97.067% of the time. ⓘ

OverviewScoring

Column impact ⓘ ↓

Search columns...

1 net\_income\_to\_total\_assets13.142%

2 persistent\_eps10.072%

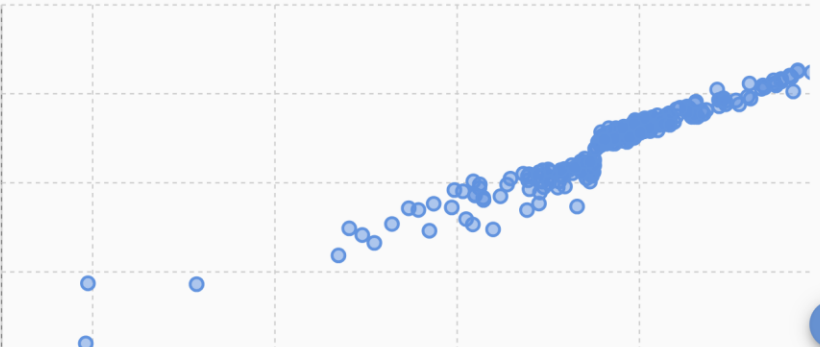
3 net\_worth\_to\_assets8.993%

4 total\_asset\_turnover

Impact of net\_income\_to\_total\_assets on prediction of bankrupt

0

Impact on prediction



bankruptcy\_training\_data(1)

Total columns: 25

Total rows: 6,819

Total cells: 170,475

bankrupt

2 category prediction

Predict



# 3b) Use model to make predictions for training data set

Amazon SageMaker

My models / bankruptcy\_prediction\_model / Version 1

+ Add version

batchInfer-bankruptcy\_prediction\_model-bankruptcy\_prediction\_data-1696522241

Prediction (bankrupt)	Probability	borrowing_d...	company	company_id	current_liabil...	current_liabil...	debt_ratio_p...
1	89.4%	0.458818609	western corp	id_6988	0.372217526	0.060766259	0.269038909
1	64.5%	0.37930429	design solutions	id_7413	0.333345174	0.041219716	0.16186474
1	51.3%	0.384998982	innocore	id_8801	0.337392013	0.060765125	0.216101823
0	83.0%	0.374219105	pharmasolve	id_9614	0.329803726	0.030201105	0.108202074
0	81.2%	0.370253398	ninetech	id_9131	0.328092756	0.021710461	0.058590561
0	80.8%	0.37450876	songster inc	id_7102	0.330409488	0.025494302	0.121292741
0	81.1%	0.374179962	rogers and sons	id_7012	0.327484654	0.047166348	0.103576503
0	78.5%	0.373113046	Hallandall ag.	id_9904	0.328042001	0.033711602	0.094385827
0	79.2%	0.377306898	Foster & Kruse	id_6905	0.331114215	0.04351399	0.144662454

Send to Amazon QuickSight

Download

# Findings from Exploratory Data Analysis

	Retained_Earnings_to_Total_Assets	Borrowing_dependency
count	6819.000000	6819.000000
mean	0.934733	0.374654
std	0.025564	0.016286
min	0.000000	0.000000
25%	0.931097	0.370168
50%	0.937672	0.372624
75%	0.944811	0.376271
max	1.000000	1.000000

From Datacorp Financials

	Net_Income_to_Total_Assets
count	6819.000000
mean	0.807760
std	0.040332
min	0.000000
25%	0.796750
50%	0.810619
75%	0.826455
max	1.000000

From Analyst A Ratios

# Recommendations

- I **recommend** including the following companies:
  - Pharmasolve, Highwood & Hart, Ninetech, Rogers and Sons, Songster inc, Foster & Kruse, and Hallaldall ag.
- I **do not** recommend the following companies:
  - Western Corp, Design Solutions, Innocore

