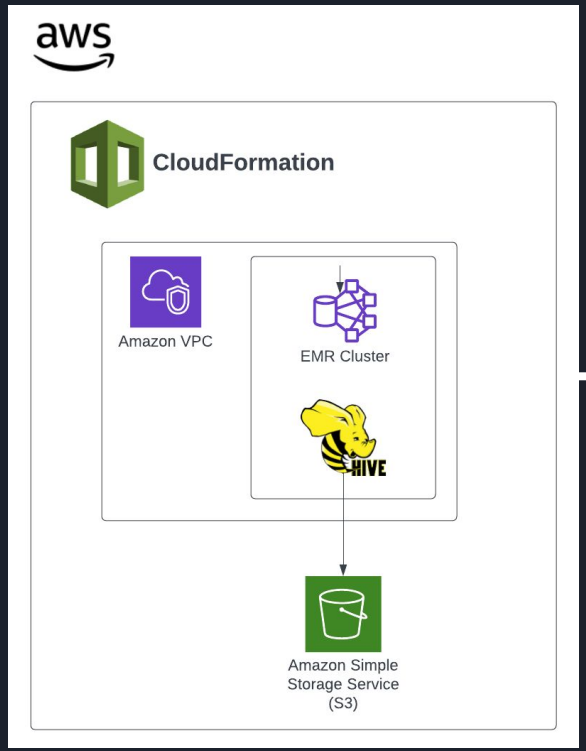
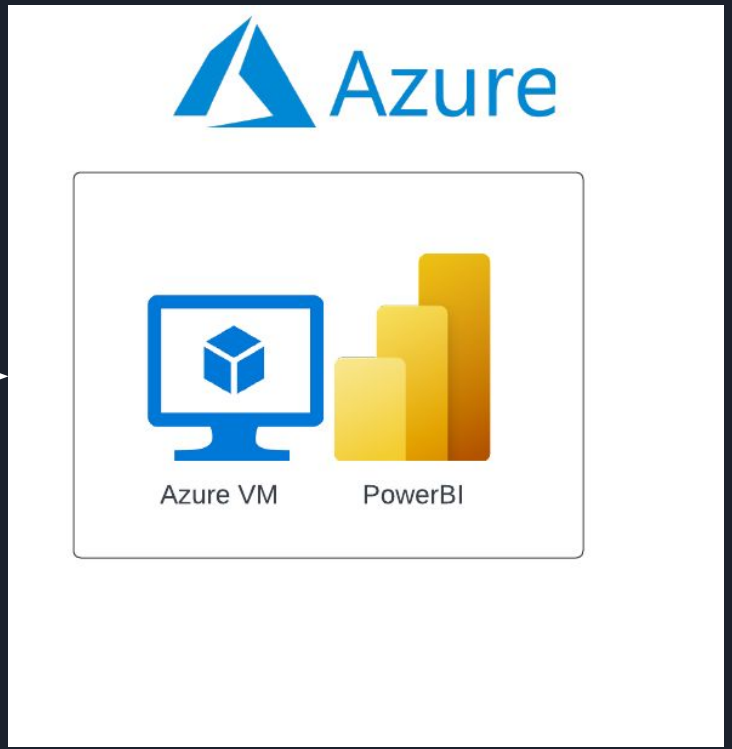
EMR Pipeline

EMR Pipeline & Data Visualization with PowerBI

Project Overview





Cloud Computing: AWS vs Azure

AWS

- Amazon
- ~60% of the Cloud Compute market
- Focus on Linux
- Well developed services and offerings
- Quicksight (data visualization) isn't great

Azure

- Microsoft
- ~20% of the Cloud Compute market
- Windows (.NET) focus
- PowerBI is a well established data visualization tool

Tools and Technologies

AWS

- Cloud9 - Web based IDE
- Cloud Development Kit (CDK)
- Virtual Private Cloud (VPC)
- Simple Storage Service (S3)
- Elastic Mapreduce (EMR)
- Apache Hive

Azure

- Virtual Machine
- Microsoft PowerBI
    - Need Windows to run PowerBI

AWS Cloud Development Kit (CDK)

- Open-source software development framework from AWS that allows developers to define cloud infrastructure as code using familiar programming languages like TypeScript, JavaScript, Python, C#, and Java.
- Developers can create infrastructure as code in the form of reusable components called constructs, which are defined using programming languages and can be easily shared with other developers.
- These constructs can then be used to generate AWS CloudFormation templates that create and manage AWS resources.
- Focus on higher level of abstraction while still providing the flexibility and control that developers need to create and manage their infrastructure.

AWS Cloud 9 Setup

- Login to Console
- Search for Cloud 9
- Create new Cloud 9 environment

UUID vs NanoID

UUID

- **Universal unique identifier**
- **You could generate 1 billion UUIDs per second for 85 and still only have a 50% chance of creating a duplicate**
- **Alphabet: 0-9, a-f**
- **36 characters**
- **Example: 06009c60-864a-4d0b-98ee-b638df53211b**
- **Size: 483 bytes**

NanoID

- Tiny, URL friendly, unique string
- 2.2 million unique ids per second
- Example: AwjDzk_GyqlPqrV2Z4OT8
- Alphabet: A-Za-z0-9_-
- 21 characters
- 60% faster than UUID
- Size: 108 bytes

AWS CDK Setup

- Open Cloud9 Instance
- Install AWS CDK using the following command:
  - python -m pip install aws-cdk-lib
- Check version to make sure it has been installed
  - cdk –version
- Create a new directory for the project
  - mkdir emr-etl-pipeline
- Create a new cdk app with the following command
  - cdk init app –language python
- Edit and install requirements
- Create a unique identifier using NanoID

Create Bucket Deployment Stack

- Get sales data and put inside data directory
- Create new folder for stack
- Create new python file for stack
- Inside the new file:
  - Create a BucketDeploymentStack class
  - Create primary bucket
  - Create log bucket
  - Create data deployment stack

Create Security Stack

- Create new folder for security stack
- Create new python file for security stack
- Create security stack class
- Create a VPC within the new class

Data Overview

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	region	country	item_type	sales_chann	order_priorit	order_date	order_id	ship_date	units_sold	unit_price	unit_cost	total_revenu	total_cost	total_profit
2	Middle East and North Africa	Libya	Cosmetics	Offline	M	10/18/2014	686800706	10-31-2014	8446	437.2	263.33	3692591.2	2224085.18	1468506.02
3	North America	Canada	Vegetables	Online	M	11-07-2011	185941302	12-08-2011	3018	154.06	90.93	464953.08	274426.74	190526.34
4	Middle East and North Africa	Libya	Baby Food	Offline	C	10/31/2016	246222341	12-09-2016	1517	255.28	159.42	387259.76	241840.14	145419.62
5	Asia	Japan	Cereal	Offline	C	04-10-2010	161442649	05-12-2010	3322	205.7	117.11	683335.4	389039.42	294295.98
6	Sub-Saharan Africa	Chad	Fruits	Offline	H	8/16/2011	645713555	8/31/2011	9845	9.33	6.92	91853.85	68127.4	23726.45
7	Europe	Armenia	Cereal	Online	H	11/24/2014	683458888	12/28/2014	9528	205.7	117.11	1959909.6	1115824.08	844085.52
8	Sub-Saharan Africa	Eritrea	Cereal	Online	H	03-04-2015	679414975	4/17/2015	2844	205.7	117.11	585010.8	333060.84	251949.96
9	Europe	Montenegro	Clothes	Offline	M	5/17/2012	208630645	6/28/2012	7299	109.28	35.84	797634.72	261596.16	536038.56
10	Central America and the Caribbean	Jamaica	Vegetables	Online	H	1/29/2015	266467225	03-07-2015	2428	154.06	90.93	374057.68	220778.04	153279.64
11	Australia and Oceania	Fiji	Vegetables	Offline	H	12/24/2013	118598544	1/19/2014	4800	154.06	90.93	739488	436464	303024
12	Sub-Saharan Africa	Togo	Clothes	Online	M	12/29/2015	451010930	1/19/2016	3012	109.28	35.84	329151.36	107950.08	221201.28
13	Europe	Montenegro	Snacks	Offline	M	2/27/2010	220003211	3/18/2010	2694	152.58	97.44	411050.52	262503.36	148547.16
14	Europe	Greece	Household	Online	C	11/17/2016	702186715	12/22/2016	1508	668.27	502.54	1007751.16	757830.32	249920.84
15	Sub-Saharan Africa	Sudan	Cosmetics	Online	C	12/20/2015	544485270	01-05-2016	4146	437.2	263.33	1812631.2	1091766.18	720865.02
16	Asia	Maldives	Fruits	Offline	L	01-08-2011	714135205	02-06-2011	7332	9.33	6.92	68407.56	50737.44	17670.12
17	Europe	Montenegro	Clothes	Offline	H	6/28/2010	448685348	7/22/2010	4820	109.28	35.84	526729.6	172748.8	353980.8
18	Europe	Estonia	Office Supplies	Online	H	4/25/2016	405997025	05-12-2016	2397	651.21	524.96	1560950.37	1258329.12	302621.25
19	North America	Greenland	Beverages	Online	M	7/27/2012	414244067	08-07-2012	2880	47.45	31.79	136656	91555.2	45100.8
20	Sub-Saharan Africa	Cape Verde	Clothes	Online	C	09-08-2014	821912801	10-03-2014	1117	109.28	35.84	122065.76	40033.28	82032.48
21	Sub-Saharan Africa	Senegal	Household	Offline	L	8/27/2012	247802054	09-08-2012	8989	668.27	502.54	6007079.03	4517332.06	1489746.97
22	Australia and Oceania	Federated State	Snacks	Online	C	09-03-2012	531023156	10/15/2012	407	152.58	97.44	62100.06	39658.08	22441.98
23	Europe	Bulgaria	Clothes	Online	L	8/27/2010	880999934	9/16/2010	6313	109.28	35.84	689884.64	226257.92	463626.72
24	Middle East and North Africa	Algeria	Personal Care	Online	H	2/20/2011	127468717	03-09-2011	9681	81.73	56.67	791228.13	548622.27	242605.86
25	Asia	Mongolia	Clothes	Online	L	12-12-2015	770478332	1/24/2016	515	109.28	35.84	56279.2	18457.6	37821.6
26	Central America and the Caribbean	Grenada	Cereal	Online	H	10/28/2012	430390107	11/13/2012	852	205.7	117.11	175256.4	99777.72	75478.68
27	Central America and the Caribbean	Grenada	Beverages	Online	M	1/30/2017	397877871	3/20/2017	9759	47.45	31.79	463064.55	310238.61	152825.94

AWS Elastic Mapreduce (EMR)

- Managed big data platform on AWS
- Allows users to easily process and analyze vast amounts of data using tools such as Apache Hadoop, Spark, and Hive.
- Used for a wide range of big data tasks, including data ingestion, processing, transformation, and analysis, as well as machine learning and Spark streaming.
- EMR is highly scalable (easy to dd or remove nodes to their cluster)
- Various security features, such as encryption for data in transit and at rest, and fine-grained access control
- EMR integrates with other AWS services, and can be easily used in conjunction with other AWS big data services, such as Amazon Redshift and Amazon Athena.
- EMR also provides various management and monitoring tools, such as Amazon CloudWatch, AWS CloudTrail, and AWS Management Console, to help users manage and monitor their big data workflows.

Hive vs Spark

Hive

- Data warehousing system for querying and analyzing large datasets stored in Hadoop Distributed File System (HDFS)
- Designed for batch processing and is optimized for long-running queries over large data sets.
- Very stable
- Easier to learn since Hive Query Language (HQL) is very similar to SQL

Spark

- Spark, on the other hand, is a fast and flexible big data processing engine
- Batch and real-time processing workloads.
- Steeper learning curve, especially coming from SQL
- Multi-language support including Python, Scala, Java

Create Hive scripts

- Create tables script
  - Create an external table for the raw data
  - Create an external table for the transformed data
- Create a transformation script
  - ETL script fixes the dates in the sales data
  - Inserts the transformed data into the new table

Create EMR Cluster Stack

- Create new folder
- Create new python file
- Create new EMR cluster stack class
- Inside the new class:
  - Create a new policy to read from the scripts directory
  - Create EMR Cluster
  - Retrieve correct subnet ID from AWS console under EC2-VPC
  - Create Hive step for creating the tables
  - Create Hive step for transforming the data

Deploying the Stacks

- We'll deploy the stacks individually
- First run cdk synth [stack id]
- Then run cdk deploy [stack id]
- Deploy Data Deployment Stack
- Deploy Security Stack
- Deploy EMR Cluster Stack