ETL DESIGN

|  |
| --- |

I have chose AWS cloud as tool in order to design ETL for given task as it more reliable , cost effective , high availability, much efficient.

**DESCRIPTION:-**

1. **AWS Lamda function service** - in order to trigger the job as we can run code without provisioning or managing infrastructure as it is servless service. Automatically respond to code execution requests at any scale, from a dozen .Save costs by paying only for the compute time you use. Most importantly it supports any language script to trigger the job.
2. **AWS Pipeline-**  is built on a distributed, highly available infrastructure designed for fault tolerant execution of your activities. If failures occur in your activity logic or data sources, AWS Data Pipeline automatically retries the activity. If the failure persists, AWS Data Pipeline sends you failure notifications via Amazon SNS. Creating a pipeline is quick and easy via our drag-and-drop console. Common preconditions are built into the service, so you don’t need to write any extra logic to use them. It provide us the Template which make our task much easier.AWS Data Pipeline allows you to take advantage of a variety of features such as scheduling, dependency tracking, and error handling. You can configure AWS Data Pipeline to take actions like run Amazon EMR jobs, execute SQL queries directly against databases, or execute custom applications running on Amazon EC2 or in your own datacenter.
3. **AWS GLUE-** AWS Glue reduces the time it takes to start analyzing your data from months to minutes. It provides you with both visual and code-based interfaces to make data preparation easy.
4. **Simple Storage Service** - Scale storage resources to meet fluctuating needs with 99.999999999% of data durability.store data acrosS Amazon S3 storage classes to reduce costs without upfront investment or hardware refresh cycles. Protect your data With unmatched security, compliance, and audit capabilities. It also has Intelligent where it helps to reduce the cost of storage.
5. **Cloud Watch** - It's an option , it can help to track where servers are working and up. And if anything fails then it would send an sms alert.

events per day to hundreds of thousands per second

**Script for ETL job**

from \_\_future\_\_ import print\_function

import boto3

import urllib

print('Loading function')

glue = boto3.client('glue')

def lambda\_handler(event, context):

gluejobname="GLUE\_JOB\_NAME"

try:

runId = glue.start\_job\_run(JobName=gluejobname)

status = glue.get\_job\_run(JobName=gluejobname, RunId=runId['JobRunId'])

print("Job Status : ", status['JobRun']['JobRunState'])

except Exception as e:

print(e)

raise e

Please Note - AWS Lambda function will automatically take care for rest while we configure, we do not require any extra SQL dialect.