22.3 Lesson Summary - Cloud ETL

Amazon Web Services (AWS) is the preeminent cloud computing service. AWS offers the ability to store data in databases or hard disk style file storage, run virtual machines to process or share data, and a number of other cloud-based services. AWS is a good fit for Cloud based Extract, Transfer and Load (ETL) because the data storage and processing options are inexpensive and accessible.

Concept: The services provided by a cloud service provider, like AWS, are often utilized in cloud based ETL. Files are stored in an AWS S3 bucket which acts like a hard drive that can be accessed from anywhere. Data is **extracted** from the stored files. The data can be **transformed** on an AWS virtual machine or Google Colabratory. The final, organized data can be stored (**loaded**) into an (Relational Database Service) RDS relational database or an Amazon DocumentDB MongoDB compatible database.

Concept: AWS' **Relational Database Service** (**RDS**) provides a SQL based relational database to store data. To create a RDS database choose the RDS option from the services dropdown. Select the "Create database" option. Specify the database options keeping in mind that larger, more performant databases can be quite expensive. Once you have created your RDS database you can query it just like a local database. You can connect to a RDS database from pgAdmin on your local machine by changing the inbound network settings on your database and copying the network connection from the your RDS database's AWS webpage.

* Activity: 02-Evr\_RDS\_CRUD
* Suppl link: https://aws.amazon.com/rds/

Concept: AWS' **Simple Storage Service** (**S3**) provides the ability to store files in the cloud as you might on a local hard disk. Files are stored in buckets, which are like folders, using key-value pairs. To create a bucket, select the S3 option from the services dropdown. Choose the Create bucket option. Name your bucket, describe how it will be accessed and specify any additional properties for your bucket. You are charged by the amount of data you store in your bucket per month, so keep this in mind. You can create folders and upload files to your bucket from the webpage for your bucket. You can make files publicly available from your S3 bucket.

* Activity: 01-Evr\_S3
* Suppl link: https://aws.amazon.com/s3/

Concept: To access S3 files from Google Colabratory you can use the following code:

*from pyspark import SparkFiles*

*url="https://<bucket name>.s3.amazonaws.com/your\_file.txt"*

*spark.sparkContext.addFile(url)*

*print(SparkFiles.get("your\_file.txt"))*

To write a Spark DataFrame to your RDS database instance you can run the following code:

*mode = "append"*

*jdbc\_url="jdbc:postgresql://<connection string>:5432/<database-name>"*

*config = {"user":"postgres",*

*"password": "<password>",*

*"driver":"org.postgresql.Driver"}*

*my\_df.write.jdbc(url=jdbc\_url, table='active\_user', mode=mode, properties=config)*

* Activity: 03-Ins\_ETL\_S3\_RDS, 04-Stu\_ETL\_S3\_Colab

Concept: You may be billed for any resources you leave open on AWS so it is a good idea to delete them if you're not using them. You can do this from your AWS dashboard by browsing to each of the different services.