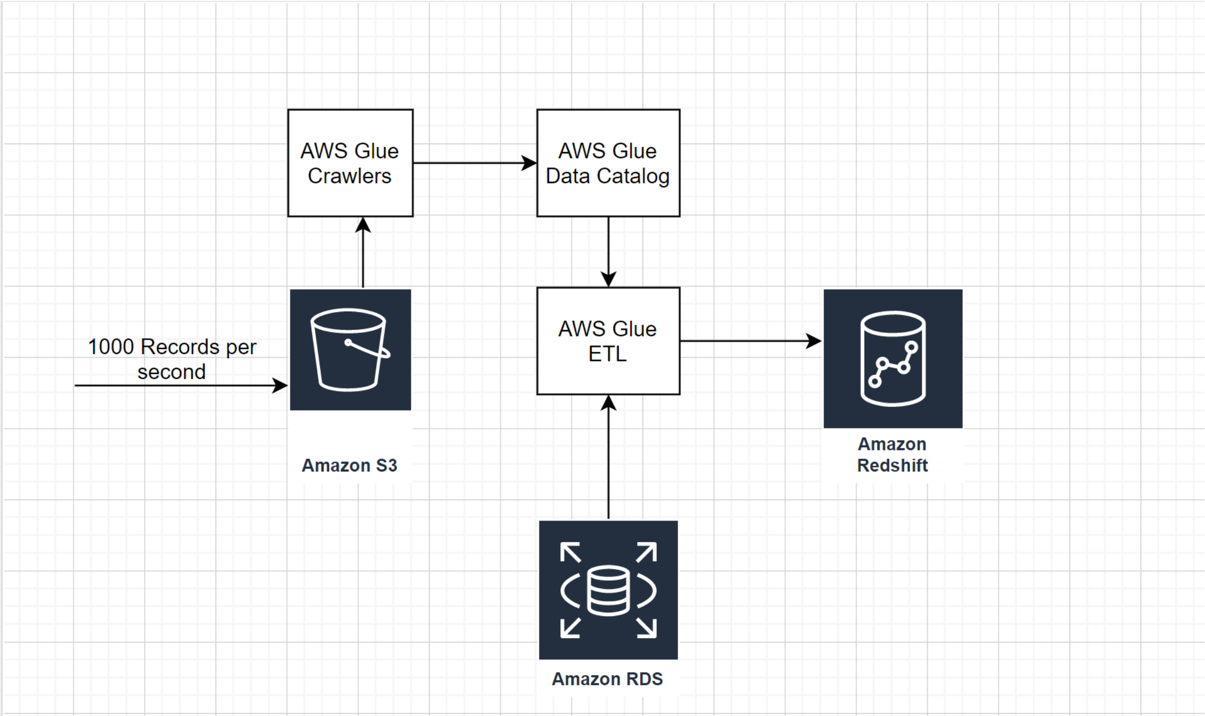
ETL Pipeline



Tool of Choice:

Amazon Web Services, the world's largest cloud computing platform, is my primary choice to host a cloud platform and build a ETL data pipeline. AWS is the most used cloud computing platform, and is better in almost all the aspects like flexibility, scalability, security, consistency when compared to its competitor's cloud platform.

Amazon S3: Since we are receiving 1000 records per second, my choice to store this data will be inside Amazon S3 buckets. Amazon S3 is a great resource to store the incoming data, and it provides with a data-driven approach to data security, management efficiency, and storage optimization. Amazon S3 also gives offers multiple option for migrating data from one source to another, making it a great resource to store large amounts of data.

AWS Glue: AWS Glue is an extract, transform and load (ETL) service offered by AWS. When AWS Glue is pointed to a data source, in this case Amazon S3, the AWS Glue Crawlers will discover the data and stores the table definition and schema in the AWS Glue Data Catalog. The Glue Data Catalog serves as a central metadata repository in the data pipeline. In this Data pipeline, AWS Glue will help in transforming the data present in Amazon S3, and will make the data ready to be loaded for analytics.

Amazon RDS: Amazon Relational Database Service enables developers to manage relational databases in cloud. Amazon RDS is compatible with commonly used database engines, and is fairly easy to implement. In the above data pipeline, I selected Amazon RDS to enrich the data further, by joining AWS Glue with a relational database. This will enhance the data quality in the AWS Glue, which will make the data ready to be loaded for data warehouse and analytics.

Amazon Redshift: Amazon Redshift is a data warehouse service offer by Amazon Web Services that is used to collect and store data. Amazon Redshift is the fastest data warehouse available, and also costs less to operate than any other cloud data warehouse. In the data pipeline above, after AWS Glue transforms the data, the resulting data is finally stored to Amazon Redshift. Amazon Redshift also enables users to analyze the data using BI tools, performing data visualization (which can also be done by Amazon QuickSight) and simplifies the process of handling large scale data sets.

Processing Frequency:

Amazon S3 has a processing frequency of at least 3500 requests per second to add data, and 5500 requests per second to retrieve data. So, in the data pipeline above which is receiving 1000 records per second, Amazon S3 is an ideal storing service.

Other tools used above in the data pipeline have a high processing frequency, and can accommodate influx of 1000 records per second into Amazon S3.

Cost:

Amazon S3: Amazon Free Tier offers free 5 GB storage, 20000 Get Requests, 2000 Put Requests for 12 months at no cost.

Amazon S3 Standard Pricing after 1 year of usage:

Amount of Data	**Cost**
First 50 TB/month	$0.023/ GB
Next 450 TB/Month	$0.022/ GB
Over 500 TB/Month	$0.021/ GB

AWS Glue:

Amazon Free Tier offers free storage for 1 million objects stored in AWS Glue Data Catalog. Additionally, 1 million requests can be made for free to the AWS Glue Data Catalog.

AWS Glue has a price of $0.44 per Data Processing Units(DPU)-hour, billed per second.

Amazon RDS: Amazon provides free uses of Amazon RDS for MySQL for up to 750 instance hours per month, with 20 GB free data base and 20 GB of backup storage for free per month. Amazon RDS is billed in one-seconds increments, and the exact price depends on the use of the developer.

Amazon Redshift:

Amazon Free Tier offers 2 months free trial of Amazon Redshift, with a limit of 750 DC2 Large Nodes hour. Amazon Redshift is the cheapest data warehouse, with a price of $0.25 per hour, with a scalability to store petabytes of data.