

Architecture Description

The project creates a pipeline to capture changes in a database and replicate those changes to another storage system. The architecture will involve several components, including a database (referred to as the "mythical database"), an RDF (Resource Description Framework) database, Amazon RDS Relational Database Service, a S3 bucket (a type of storage service provided by Amazon Web Services), and a Data Migration Service (DMS).

The first step in the architecture is to load all the existing data from the RDS into the S3 bucket. This is referred to as the "full load." Once this is complete, the pipeline will capture any changes that occur in the RDS using a process called change data capture (CDC). The changes will be written to the RDS database and Temporary HDFS / S3 Storage which acts as a buffer between the RDS and the S3 bucket.

The DMS is responsible for moving data between the RDS database and the S3 bucket. It has two endpoints: one for reading data from the RDS database and one for writing data to the S3 bucket. The DMS will periodically read data from the RDS database and write it to the S3 bucket. This process will capture any changes that have occurred in the mythical database since the last time the DMS ran.

The S3 bucket is the final destination for the replicated data. The architecture includes a temporary bucket to hold data during the replication process. Once the data has been successfully replicated to the S3 bucket, it can be used for analysis or other purposes.

Overall, the architecture appears to be designed to capture changes in a database and replicate those changes to another storage system in a reliable and efficient way. The use of a buffer (the RDF database) and a data migration service (DMF) helps to ensure that the replication process is robust and can handle large amounts of data.