API  
[MergedFile (snl.int)](http://miodataprd02.snl.int/SNL.Services.Data.Api.Service/v2/Internal/Editable/DocsFileVersions(A7841097-1CE2-423B-88C2-7D6F0B7D7598)/$value)

Database mysql(UDR)

KVF (key file version ID )



Glue Job  
Apache Tika (python library)

* Data filter
* Extract paragraphs
* Cleaning and filtering the data

Elastic Search

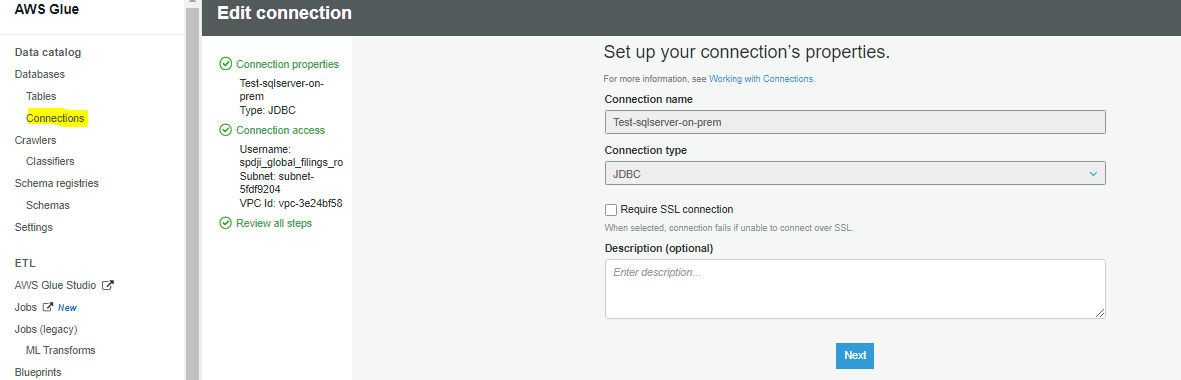
* Storing the data
* Filtering data

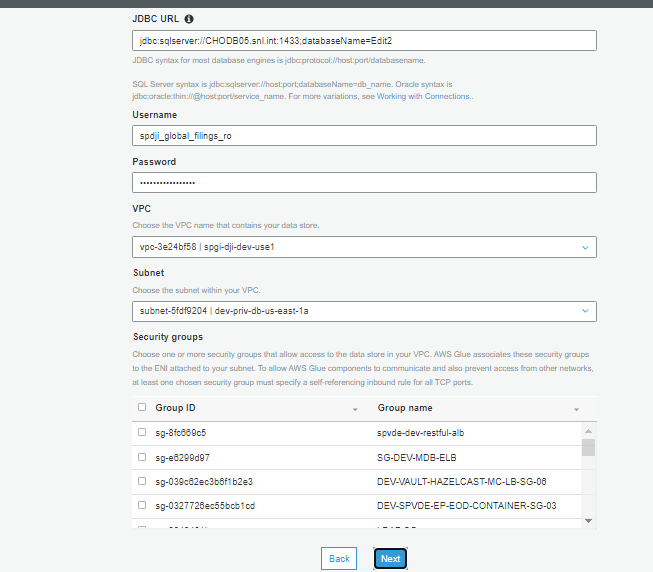
S3

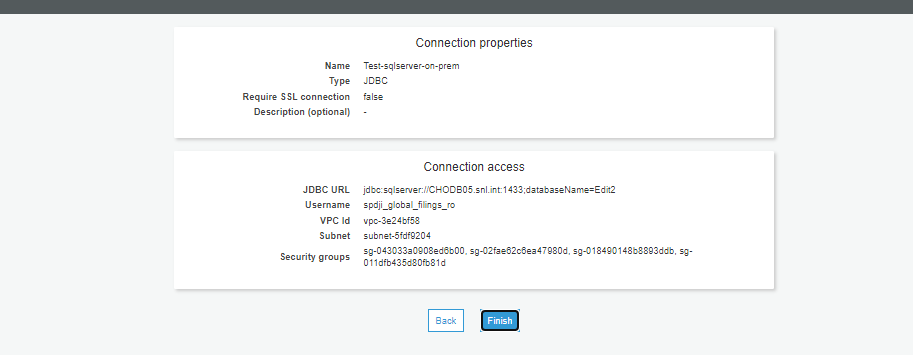
* Models (condition are mentioned in the models
* Models will give us confirmation is the condition is Correct or not

Final data set are stored in  
  
Mysql serverless Database

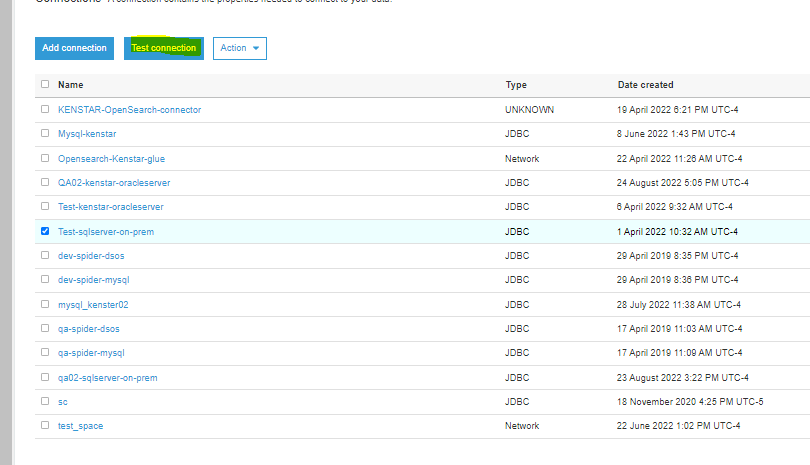
Creating the Glue connection

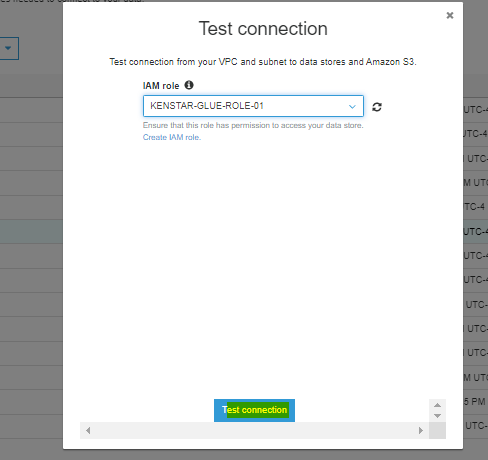


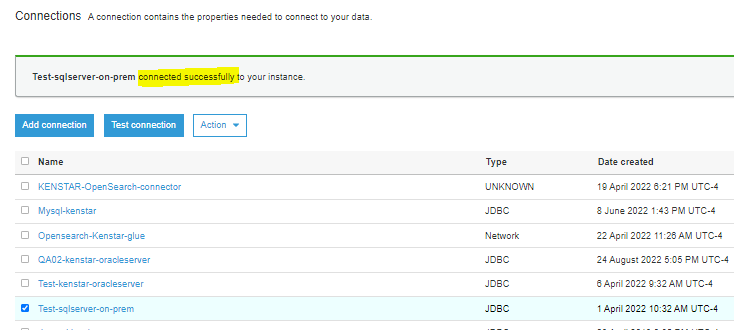




Testing Glue Connection:







Creating Glue Job:

Create AWS Glue job: Name –> qa-kenstar2-etl

* **Basic Properties**
  + - IAM Role 🡪 KENSTAR-GLUE-ROLE-01
    - Type: 🡪Spark
    - Glue Version🡪 Glue 3.0 – Supports spark 3.1, scala 2, python 3
    - Language🡪 Python 3
    - Worker Type🡪 G 2X
    - Automatically Scale the number of workers 🡪 Enable
    - Maximum Number of workers 🡪 50
    - Checked Generate job insights
    - Job Bookmark 🡪 Enable
    - Number of retires 🡪 1
    - Job Time out 🡪 2880
* **Advance properties**
  + - Script Filename 🡪 etl-global.py
    - Script Path 🡪 s3://qa-kenstar-01/ETL/
    - Job Metrics 🡪 Enable
    - Continuous logging 🡪 Enable
    - Spark UI 🡪 Enable
    - Spark UI Log path 🡪 s3://qa-kenstar-01/ETL/Glue/
    - Maximum Concurrency 🡪 1
    - Temporary Path 🡪 s3://qa-kenstar-01/ETL/Glue/
    - Specify configuration 🡪 None
    - Use data catalog as the Hive meta store 🡪 Enable

