Files
main
Go to file t
.gitignore
LICENSE
README.md
create_sample_data.py
etl-phase1.md
etl-phase1.png
etl-config.template
field_map.txt
final-phase.md
final-phase.png
hl7_field_names_to_ignore.txt
license.txt
redis_json_example.png
redis_json_key_name_sample.png
requirements.txt
s3_json_to_psql_etl.py
s3_redis_json_to_psql_etl.py
sample.json

DataTooling / README.md
History

Preview Code Blame 245 lines (197 loc) · 13 KB Raw

ETL Script processes HL7 data from JSON files in S3 buckets or RedisJSON and stores the transformed data in PostgreSQL schema.

See LinkedIn Post

Requires:

```
Use create_sample_data.py to generate anonymized sample data

1. Apache Spark Standalone or Cluster (3.4.1 or later)
   I set up a Standalone Spark Master with 5 worker nodes
   https://spark.apache.org/downloads.html
2. Have following JARS downloaded - in my case for Spark v3.4.1
   /var/lib/spark/jars/postgresql-42.6.0.jar
   /var/lib/spark/jars/aws-java-sdk-bundle-1.12.262.jar
   /var/lib/spark/jars/hadoop-aws-3.3.4.jar
3. Postgres JDBC driver (postgresql-42.6.0.jar or later)
4. S3 Bucket name (s3_bucket_name) (Multi connect is using
                                    RAW HL7 messages as JSON files
                                    in a S3 bucket)

   5. S3 Bucket Prefix (yyyy/m/d)
   6. S3 Bucket contains, single, or a set of JSON files that are
      formatted like the Sample JSON file.
   7. etl.config file to store configuration details etc:
      [reportdb]
      host=
      port=5432
      dbname=
      dbuser=
      dbuserpass=

      [spark]
      master=
      masterport=7077

      [redis]
      host=
      port=6379

      [aws]
      access_key=
      secret_key=
      [mappings]
      IGNORE_GEO_FIELDS=('FID_1','FID_12','FV1_1','IN1_1','EVN_1'
      IGNORE_COMPONENT_FIELDS=('OC_1','ZL_5','XTN_2','XTN_3','XTN
      HL7_SEGMENTS=('pid','pv1','pv2','aud','exe','ipd','ipd','ob
   8. List of fields to ignore stored in "hl7_field_names_to_ignor
      Sample:
         alt_l_set_id_kli_ul_name_1
         evn_1_event_type_code_id_name
         evn_5_operator_id_xcn_2_family_name_2
         evn_5_operator_id_xcn_3_given_name_2
         evn_5_operator_id_xcn_5_assigning_authority_1
         evn_5_operator_id_xcn_6_name_type_1
         gt1_12_guarantor_relationship_ce_1_identifier_1
         gt1_12_guarantor_xsn_st_name_1
         gt1_16_guarantor_employer_address_xad_1_family_name_1
         gt1_17_guarantor_employer_address_xad_1_street_address_
         gt1_17_guarantor_employer_address_xad_5_state_or_provin
         gt1_17_guarantor_employer_address_xad_9_zip_or_postal_c
         gt1_18_guarantor_employer_phone_number_xtn_1_telephone_
         gt1_20_guarantor_employment_status_is_name_1
         gt1_8_guarantor_relationship_ce_1_identifier_1
         gt1_2_guarantor_number_cx_1_id_number_1
         gt1_3_guarantor_name_xpn_5_suffix_e_g_jr_or_iii_1
         gt1_3_guarantor_name_xpn_3_prefix_e_g_dr_1
         gt1_3_guarantor_name_xpn_7_name_type_code_1
         gt1_3_guarantor_name_xpn_1_family_name_1
         gt1_4_guarantor_spouse_name_xpn_2_given_name_1
         gt1_5_guarantor_address_xad_1_street_address_1
      Field mapping file "field_map.txt",
         field_name = f"{ar_name}_{ke_long_name}_{fn_name}_{fc_l
         Sample:
            {'pid_2_patient_id_cx_1_id_number': 'pt_id_k'}
            {'pid_38_patient_death_indicator_id_name':'pt_death_ind
            {'pid_3_patient_identifier_list_cx_1_id_number_2':'pt_i
            {'pid_3_patient_identifier_list_cx_1_id_number_3':'pt_i
            {'pid_4_alternate_patient_id_pid_cx_1_id_number':'pt_id
            {'pid_5_patient_name_xpn_1_family_name':'pt_last_name'}
```

How-to Run - S3:

```
./s3_json_to_psql_etl.py --add-feed-name "acme,roadrunner" --s

23/09/02 16:15:00 WARN NativeCodeLoader: Unable to load native-
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For Spark

**** Starting for acme

23/09/02 16:15:10 WARN MetricsConfig: Cannot locate configurati
23/09/02 16:15:53 WARN package: Truncated the string representa

**** Stored data in table v4_acme
**** Completed for acme
**** Starting for roadrunner
**** Stored data in table v4_roadrunner
**** Completed for roadrunner
```

How-to Run - RedisJSON:

```
./s3_redis_json_to_psql_etl.py --add-feed-name redis_$(date '+%

23/10/05 18:22:37 WARN NativeCodeLoader: Unable to load native-
Setting default log level to "WARN".

**** Starting for redis_1696626545
**** Get Redis Keys ****
**** Lower Case Column Names ****
**** Process HL7 ****
**** Start Process ****
```

Example RedisJSON:

Sample RedisJSON Key Names:

Total: 1.010.555

```
JSON    pid_8537764_d6c713a105814f9aabc3fb399ba31f22.json

JSON    pid_7260099_ffa85a85808249638e1cb5a9bab12a15.json

JSON    pid_3762940_d3bc3432552b4dd7bb6f7f0e4fc4d3fe9.json
```

Sample RedisJSON Key Content:

```
JSON    pid_8537764_d6c713a105814f9aabc3fb399ba31f22.json
Key Size 6018   Length 6015   TTL: No limit                    Last refresh: 11:00:
{"_id":"6537f0ff4a7e7c7c7c7c7c7c","fileformatname":"HL7","time-to-live":"10","BOTTOM":"qwertyuiop","DATE":
"2 17:07:","YOUNG":"2023-09-18 17:07:","YOUNGNAME":"ZZZZZZZZ","YOUNGS":
"MSH|^~\&|CLOVER|CLOVER|||20230918170700||ADT^A08|P-00001|P|2.5|||||||
~EVN|A08|20230918170700||||20230918170700~PID|1||P-00001||~PV1|1|O|||||||
~OBX|1|TS|^^^|1|20230918170700|||||||~OBX|2|TS|^^^|2|20230918170700|||||||
~OBX|3|TS|^^^|3|20230918170700|||||||~OBX|4|TS|^^^|4|20230918170700|||||||
~OBX|5|TS|^^^|5|20230918170700|||||||~OBX|6|TS|^^^|6|20230918170700|||||||
~OBX|7|TS|^^^|7|20230918170700|||||||~OBX|8|TS|^^^|8|20230918170700|||||||
~OBX|9|TS|^^^|9|20230918170700|||||||~OBX|10|TS|^^^|10|20230918170700|||||||
~OBX|11|TS|^^^|11|20230918170700|||||||~OBX|12|TS|^^^|12|20230918170700|||||||
~OBX|13|TS|^^^|13|20230918170700|||||||~OBX|14|TS|^^^|14|20230918170700|||||||
~OBX|15|TS|^^^|15|20230918170700|||||||~OBX|16|TS|^^^|16|20230918170700|||||||
~OBX|17|TS|^^^|17|20230918170700|||||||~OBX|18|TS|^^^|18|20230918170700|||||||
~OBX|19|TS|^^^|19|20230918170700|||||||~OBX|20|TS|^^^|20|20230918170700|||||||
~OBX|21|TS|^^^|21|20230918170700|||||||~OBX|22|TS|^^^|22|20230918170700|||||||
~OBX|23|TS|^^^|23|20230918170700|||||||~OBX|24|TS|^^^|24|20230918170700|||||||
~OBX|25|TS|^^^|25|20230918170700|||||||~OBX|26|TS|^^^|26|20230918170700|||||||
~OBX|27|TS|^^^|27|20230918170700|||||||~OBX|28|TS|^^^|28|20230918170700|||||||
~OBX|29|TS|^^^|29|20230918170700|||||||~OBX|30|TS|^^^|30|20230918170700|||||||
~OBX|31|TS|^^^|31|20230918170700|||||||~OBX|32|TS|^^^|32|20230918170700|||||||
~OBX|33|TS|^^^|33|20230918170700|||||||~OBX|34|TS|^^^|34|20230918170700|||||||
~OBX|35|TS|^^^|35|20230918170700|||||||~OBX|36|TS|^^^|36|20230918170700|||||||
~OBX|37|TS|^^^|37|20230918170700|||||||~OBX|38|TS|^^^|38|20230918170700|||||||
~OBX|39|TS|^^^|39|20230918170700|||||||~OBX|40|TS|^^^|40|20230918170700|||||||
~OBX|41|TS|^^^|41|20230918170700|||||||~OBX|42|TS|^^^|42|20230918170700|||||||
~OBX|43|TS|^^^|43|20230918170700|||||||~OBX|44|TS|^^^|44|20230918170700|||||||
~OBX|45|TS|^^^|45|20230918170700|||||||~OBX|46|TS|^^^|46|20230918170700|||||||
~OBX|47|TS|^^^|47|20230918170700|||||||~OBX|48|TS|^^^|48|20230918170700|||||||
~OBX|49|TS|^^^|49|20230918170700|||||||~OBX|50|TS|^^^|50|20230918170700|||||||
~OBX|51|TS|^^^|51|20230918170700|||||||~OBX|52|TS|^^^|52|20230918170700|||||||
~OBX|53|TS|^^^|53|20230918170700|||||||~OBX|54|TS|^^^|54|20230918170700|||||||
~OBX|55|TS|^^^|55|20230918170700|||||||~OBX|56|TS|^^^|56|20230918170700|||||||
~OBX|57|TS|^^^|57|20230918170700|||||||~OBX|58|TS|^^^|58|20230918170700|||||||
~OBX|59|TS|^^^|59|20230918170700|||||||~OBX|60|TS|^^^|60|20230918170700|||||||
~OBX|61|TS|^^^|61|20230918170700|||||||~OBX|62|TS|^^^|62|20230918170700|||||||
~OBX|63|TS|^^^|63|20230918170700|||||||~OBX|64|TS|^^^|64|20230918170700|||||||
~OBX|65|TS|^^^|65|20230918170700|||||||~OBX|66|TS|^^^|66|20230918170700|||||||
~OBX|67|TS|^^^|67|20230918170700|||||||~OBX|68|TS|^^^|68|20230918170700|||||||
~OBX|69|TS|^^^|69|20230918170700|||||||~OBX|70|TS|^^^|70|20230918170700|||||||
~OBX|71|TS|^^^|71|20230918170700|||||||~OBX|72|TS|^^^|72|20230918170700|||||||
~OBX|73|TS|^^^|73|20230918170700|||||||~OBX|74|TS|^^^|74|20230918170700|||||||
~OBX|75|TS|^^^|75|20230918170700|||||||~OBX|76|TS|^^^|76|20230918170700|||||||
~OBX|77|TS|^^^|77|20230918170700|||||||~OBX|78|TS|^^^|78|20230918170700|||||||
~OBX|79|TS|^^^|79|20230918170700|||||||~OBX|80|TS|^^^|80|20230918170700|||||||
```