**Project Pipeline**

  

  

 

**Ingesting file from log server to S3 through Kafka:**

We are using kafka connects. Kafka Connect is a framework for connecting Kafka with external systems. It is having 2 components source connectors and sink connectors.  
**JSON Source Connector:** The Kafka Connect JSON Source connector is used to [stream JSON files](https://en.wikipedia.org/wiki/JSON_Streaming) from an SFTP directory while also converting the data based on the schema supplied in the configuration. If the schema is not supplied in the configuration then the connector can also auto generate the key.schema and value.schema at run time when schema.generation.enabled is true.

connector.class**=**io.confluent.connect.sftp.SftpJsonSourceConnector

Other Connectors are:

connector.class=io.confluent.connect.sftp.SftpCsvSourceConnector

value.converter=org.apache.kafka.connect.converters.ByteArrayConverter

We need to create properties file at */etc/kafka/connect-file-source.properties* like below:

name**=** JsonSftp

tasks.max**=**1

connector.class**=**io.confluent.connect.sftp.SftpJsonSourceConnector

input.path**=**/path/to/data

error.path**=**/path/to/error

finished.path**=**/path/to/finished

cleanup.policy**=**NONE

input.file.pattern**=**json-sftp-source.json

behavior.on.error**=**IGNORE

sftp.username**=**username

sftp.password**=**password

sftp.host**=**<IP of FTP server>

sftp.port**=**22

kafka.topic**=**sftp-json-topic

key.schema**=**{\"name\" : \"com.example.users.UserKey\",\"type\" : \"STRUCT\",\"isOptional\" : false,\"fieldSchemas\" : {\"id\" : {\"type\" : \"INT64\",\"isOptional\" : false}}}

value.schema**=**{\"name\" : \"com.example.users.User\",\"type\" : \"STRUCT\",\"isOptional\" : false,\"fieldSchemas\" : {\"id\" : {\"type\" : \"INT64\",\"isOptional\" : false},\"first\_name\" : {\"type\" : \"STRING\",\"isOptional\" : true},\"last\_name\" : {\"type\" : \"STRING\",\"isOptional\" : true},\"email\" : {\"type\" : \"STRING\",\"isOptional\" : true},\"gender\" : {\"type\" : \"STRING\",\"isOptional\" : true},\"ip\_address\" : {\"type\" : \"STRING\",\"isOptional\" : true},\"last\_login\" : {\"type\" : \"STRING\",\"isOptional\" : true},\"account\_balance\" : {\"name\" : \"org.apache.kafka.connect.data.Decimal\",\"type\" : \"BYTES\",\"version\" : 1,\"parameters\" : {\"scale\" : \"2\"},\"isOptional\" : true},\"country\" : {\"type\" : \"STRING\",\"isOptional\" : true},\"favorite\_color\" : {\"type\" : \"STRING\",\"isOptional\" : true}}}

**Sink Connector:** The Kafka Connect S3 Sink connector is used to [stream JSON files](https://en.wikipedia.org/wiki/JSON_Streaming) from a topic to S3.

We need to create properties file at */etc/kafka/connect-file-source.properties* like below:

name**=**S3Sink

tasks.max**=**1

connector.class**=** io.confluent.connect.s3.s3sinkconnector

cleanup.policy**=**NONE

behavior.on.error**=**IGNORE

kafka.topic=sftp-json-topic  
aws.s3.bucket=<s3-bucket-name>

aws.accessKeyId=<aws-access-key>

aws.secretKey=<aws-secret-key>

value.converter=org.apache.kafka.connect.converters.ByteArrayConverter

key.converter=org.apache.kafka.connect.converters.ByteArrayConverter