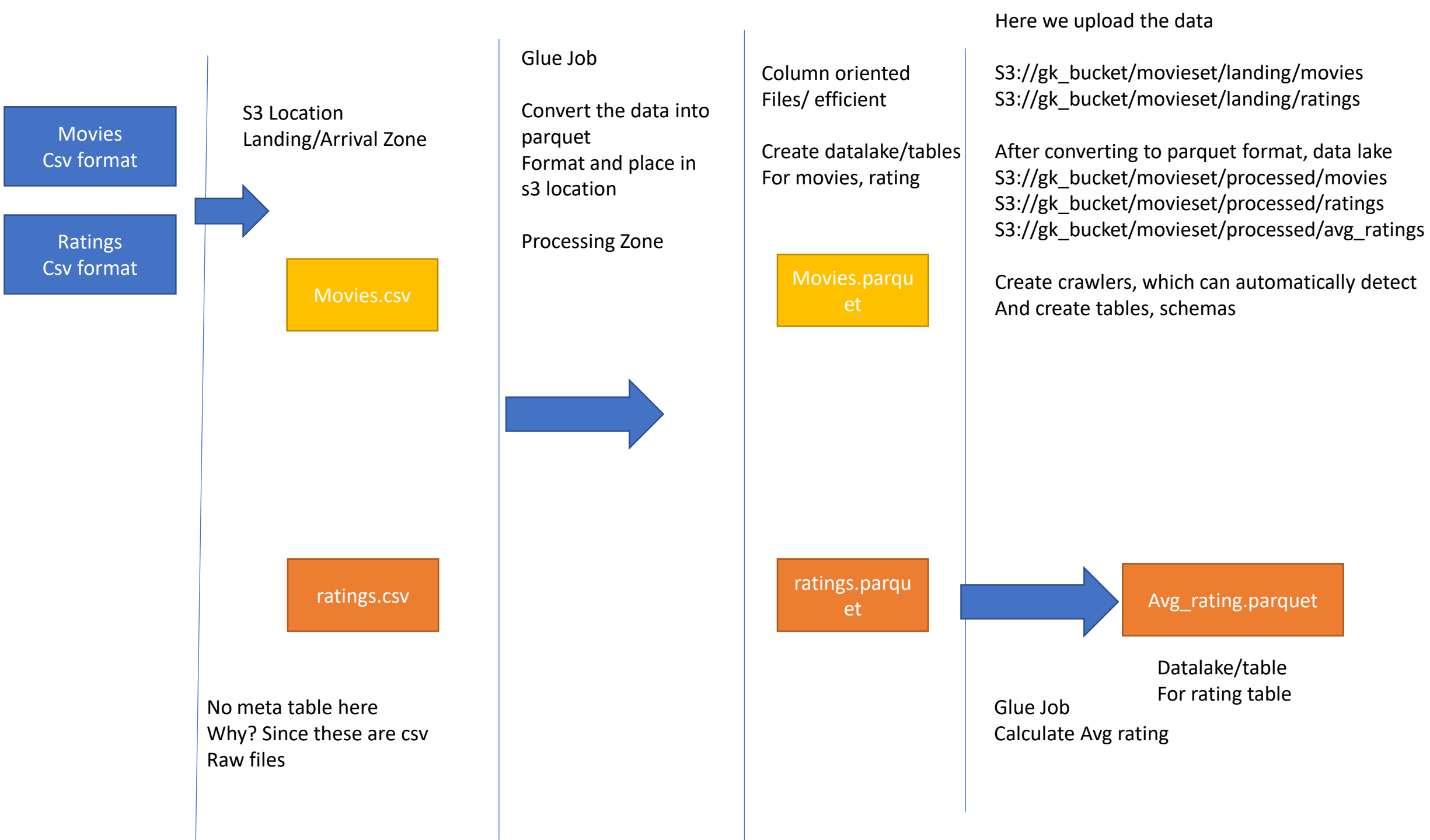
Here we upload the data
Movies
Csv format
Ratings
Csv format
S3 Location
Landing/Arrival Zone
Movies.csv
ratings.csv
No meta table here
Why? Since these are csv
Raw files
Glue Job
Convert the data into
parquet
Format and place in
s3 location
Processing Zone
Column oriented
Files/ efficient
Create datalake/tables
For movies, rating
Movies.parqu
et
ratings.parqu
et
S3://gk_bucket/movieset/landing/movies
S3://gk_bucket/movieset/landing/ratings
After converting to parquet format, data lake
S3://gk_bucket/movieset/processed/movies
S3://gk_bucket/movieset/processed/ratings
S3://gk_bucket/movieset/processed/avg_ratings
Create crawlers, which can automatically detect
And create tables, schemas
Avg_rating.parquet
Datalake/table
For rating table
Glue Job
Calculate Avg rating