**Question1**

Spray the **Movies.csv** file to the path ~yourinitial::movielens::Movies.csv

|  |  |  |
| --- | --- | --- |
| **Column\_Name** | **Data\_type** | **Description** |
| movieId | Integer | Unique Identifier for Movie |
| Title | String | Title of moview with Year of release |
| Genres | String | List of genres which movie belongs to ,Multiple genres seperated by '|' |

Read the above sprayed file as **RawMovies** and create the dataset **StagingMovies** with below information.

**MovieId :** movieid from RawMovies

**Title** (Only Title Name from RawMovies should be set here, For eg “Heat (1995)” only “Heat” should be updated

**Year** (Update year in this column which is split from RawMovies.title , for eg “Heat (1995)” “1995 should be updated”

**Genres** RawMovies.Genres

Output the file **StagingMovies** in the path ~yourinitial::movielens:: **StagingMovies**

**Question 2**

Spray the **Ratings.Csv** file to the path yourinitial::movielens::Ratings.csv

|  |  |  |
| --- | --- | --- |
| **Column\_Name** | **Data\_type** | **Description** |
| userId | Integer | Unique Identifier for user |
| movieId | String | Foriegn key from Movies table |
| rating | Decimal | Rating given by a user for a particular movie |
| timestamp | Integer | Unix Time format |

Read the above sprayed file as **RawRatings** and create the dataset **StagingRatings** with below information.

**userid :** userid from RawRatings.

**MovieID** :.movieid from RawRatings .

**Rating** := rating from RawRatings

**Timestamp** :=Convert the unix time stamp to SystemDate format and update here .

Output the file **StagingRatings** in the path ~yourinitial::movielens:: StagingRatings

**Question 3**

From the above entities Please try to answer the following questions.

* Number of movies for which belongs to the genre drama.
* Max and Minimum rating for each movie.
* Most rated movie.
* Movie which is having maximum average rating.
* Year with most number of movies released.
* User who has Rated more no of Movies and what is that Users Avg. Rating ?

Hello Xiaoming,

This is Girikratna, I am a part of the HPCC Internship Program and my project is "[Investigate Third Party Environments Working with the HPCC Systems](https://wiki.hpccsystems.com/display/hpcc/Investigate+Third+Party+Environments+Working+with+the+HPCC+Systems)" based on our previous call I have worked on and completed the Integration of Power BI with HPCC Systems wherein a SOAP Request is sent from Power BI to WsSQL and the output is in tabular form on Power BI, Initially it was planned that I work on the Azure Synapse Analytics Integration but due to Firewall and Security concerns that was put on hold.

I am planning an Integration of GitPod with HPCC Systems wherein GitPod can be used as simple no setup option for newer developers and trainees as it can be configured on GitPod. Just for an Introduction, GitPod works by provisioning a VM to every Workspace with an IDE on the Browser, it can all be configured via a yaml file.

Enabling this with Github as native support, any new developer or trainee can access and work on codebases on GitHub without the need to set-up and install any IDE, for the future a web based ECL IDE can be configured via the OIDC (as part of the GitPod Documentation) and directly run on the browser.

Do let me know your thoughts on this technology and the specified use case, let me know if you have any other use cases and suggestions in mind.

Thanks and Regards,

Girikratna Sharma

Trainee Software Engineer  
girikratna.sharma@lexisnexisrisk.com  
LexisNexis Risk Solutions – Mumbai