Our data is sourced from 4 IMDB datasets from the IMDB website. Dataset details sourced from IMDB Link

Data Ingestion:

The getdata.sh script gets the dataset from the link : <https://www.imdb.com/interfaces/>

The data is first available in .tsv format and is in a compressed format, we process these .tsv files as temporary tables and then subsequently hive managed tables of ORC format. This is done separately for all the 4 files in [IMDB\_crew.hql](https://drive.google.com/file/d/1u6VLxWdbXMu9fV6kFQxMtS2CCStI0ZHa/view?usp=share_link), [IMDB\_name.hql](https://drive.google.com/file/d/1vTQtTFs2EjieT1MgLIm1TnHuRoWMX50u/view?usp=share_link), [IMDB\_rating.hql](https://drive.google.com/file/d/1fMfeNdws7e4whkvjNGIiU5M1p7F-LoLw/view?usp=share_link), [IMDB\_title.hql](https://drive.google.com/file/d/1eMHmKV3h5MwJBnQQmHvkXojx3_kdvapT/view?usp=share_link)

To get data into hive, you need to first use beeline command:

**beeline -u jdbc:hive2://localhost:10000/default -n hadoop -d org.apache.hive.jdbc.HiveDriver**

Now that we’ve created hive tables, let’s move to the next step!