**OBJECTIVE :**

**This dataset describes rating activity from a movie recommendation service containing 20000263 ratings across 27278 movies. The dataset was created by 138493 users between 1995 and 2015. The dataset is spread in 2 files, movies.csv and ratings.csv. Return the top 100 popular movies with count using Python's dictionary and Spark's broadcast variables**

**# PREPARING RDDS:**

ratingsrdd = sc.textFile("/user/cdh7bmanipal26/spark\_assignment/ratings.csv")

moviesrdd = sc.textFile("/user/cdh7bmanipal26/spark\_assignment/movies.csv")

**#REMOVING HEADERS:**

header = ratingsrdd.first()

ratingrdd\_clear = ratingsrdd.filter(lambda x: x!= header)

ratingrdd\_clear.take(2)

[u'1,2,3.5,1112486027', u'1,29,3.5,1112484676']

**#FILTERING THE RATING WITH THRESHOLD OF 3.5 RATING AND FORMING A NEW DATAFRAME BASED ON THRESHOLD.**

filteredrating = ratingrdd\_clear.filter(lambda x: float(x.split(",")[2])>=3.5)

filteredrating.take(5)

5[u'1,2,3.5,1112486027', u'1,29,3.5,1112484676', u'1,32,3.5,1112484819', u'1,47,3.5,1112484727', u'1,50,3.5,1112484580']

**#FORMING A (KEY, VALUE) PAIR OF (MOVIE ID, COUNT OF RATING):**

rating =filteredrating.map(lambda x: (x.split(",")[1],1)).reduceByKey(lambda a,b: a+b).map(lambda x: (x[1], x[0])).sortByKey(ascending = False)

rating.take(10)

[(58207, u'318'), (55266, u'296'), (54022, u'593'), (51707, u'356'), (45572, u'260'), (44046, u'527'), (44020, u'2571'), (42191, u'110'), (42105, u'50'), (39269, u'589')]

rating\_count = rating.map(lambda x: (x[1], x[0]))

rating\_count.take(10)

[(u'318', 58207), (u'296', 55266), (u'593', 54022), (u'356', 51707), (u'260', 45572), (u'527', 44046), (u'2571', 44020), (u'110', 42191), (u'50', 42105), (u'589', 39269)]

**#MAKING KEY VALUE PAIR OF MOVIE ID AND ITS NAME :**

moviesrdd.first()

u'movieId,title,genres'

movieid\_title= moviesrdd.map(lambda x :(x.split(",")[0], x.split(",")[1]))

**#JOING THE MOVIE AND ITS TITLE KEY VALUE PAIR FROM ABOVE TO MOVIE ID AND RATING COUNT KEY VALUE PAIR BY USING .JOIN() FUNCTION:**

pairedrdd = movieid\_title.join(rating\_count)

pairedrdd.take(10)

[(u'74701', (u'Of Time and the City (2008)', 6)), (u'127877', (u'Kawa (2010)', 1)), (u'73399', (u'Last Resort (2000)', 3)), (u'124348', (u'The Land Before Time XIII: The Wisdom of Friends (2007)', 1)), (u'80487', (u'"Married Woman', 6)), (u'107743', (u'"Class of 92', 4)), (u'81347', (u'Vivere (2007)', 4)), (u'86368', (u'Confessions of a Nazi Spy (1939)', 1)), (u'4022', (u'Cast Away (2000)', 12932)), (u'83330', (u'Rasputin (1996)', 5))]

**#SORTING ABOVE OUTPUT WITH DESCENDING ORDER BASED ON VALUES**

>>> pairerddsorted = pairedrdd.values()

>>> pairerddsorted.take(5)

[(u'Of Time and the City (2008)', 6), (u'Kawa (2010)', 1), (u'Last Resort (2000)', 3), (u'The Land Before Time XIII: The Wisdom of Friends (2007)', 1), (u'"Married Woman', 6)]

pairerddsorted = pairedrdd.values().sortBy(lambda x: x[1], ascending =False)

**#PRINTING TOP 100 MOST POPULAR MOVIES:**

for i in pairerddsorted.take(100):print(i)

(u'"Shawshank Redemption', 58207)

(u'Pulp Fiction (1994)', 55266)

(u'"Silence of the Lambs', 54022)

(u'Forrest Gump (1994)', 51707)

(u'Star Wars: Episode IV - A New Hope (1977)', 45572)

(u"Schindler's List (1993)", 44046)

(u'"Matrix', 44020)

(u'Braveheart (1995)', 42191)

(u'"Usual Suspects', 42105)

(u'Terminator 2: Judgment Day (1991)', 39269)

(u'"Fugitive', 38943)

(u'American Beauty (1999)', 38399)

(u'Star Wars: Episode V - The Empire Strikes Back (1980)', 38350)

(u'Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)', 37944)

(u'Jurassic Park (1993)', 37849)

(u'Toy Story (1995)', 37494)

(u'"Godfather', 36742)

(u'Star Wars: Episode VI - Return of the Jedi (1983)', 36647)

(u'Fargo (1996)', 35449)

(u'Fight Club (1999)', 35139)

(u'Seven (a.k.a. Se7en) (1995)', 35056)

(u'Apollo 13 (1995)', 34705)

(u'Twelve Monkeys (a.k.a. 12 Monkeys) (1995)', 33414)

(u'"Sixth Sense', 32801)

(u'Back to the Future (1985)', 31955)

(u'"Lord of the Rings: The Fellowship of the Ring', 31935)

(u'Saving Private Ryan (1998)', 30454)

(u'"Lord of the Rings: The Two Towers', 28839)

(u'Dances with Wolves (1990)', 28685)

(u'Monty Python and the Holy Grail (1975)', 28368)

(u'"Princess Bride', 27415)

(u'Memento (2000)', 27163)

(u'"Lord of the Rings: The Return of the King', 27132)

(u"One Flew Over the Cuckoo's Nest (1975)", 26685)

(u'"Lion King', 26475)

(u'Aladdin (1992)', 26415)

(u'Gladiator (2000)', 26237)

(u'"Terminator', 25768)

(u'Indiana Jones and the Last Crusade (1989)', 25460)

(u'Alien (1979)', 25267)

(u'Blade Runner (1982)', 25145)

(u'Shrek (2001)', 25129)

(u'Independence Day (a.k.a. ID4) (1996)', 24429)

(u'Groundhog Day (1993)', 24244)

(u'Die Hard (1988)', 23905)

(u'"Godfather: Part II', 23804)

(u'Good Will Hunting (1997)', 23499)

(u'Reservoir Dogs (1992)', 23468)

(u'True Lies (1994)', 23447)

(u'Speed (1994)', 23395)

(u'Goodfellas (1990)', 22855)

(u'Men in Black (a.k.a. MIB) (1997)', 22692)

(u'E.T. the Extra-Terrestrial (1982)', 22550)

(u'L.A. Confidential (1997)', 22372)

(u'Batman (1989)', 22114)

(u'Aliens (1986)', 21718)

(u'Being John Malkovich (1999)', 21647)

(u'Beauty and the Beast (1991)', 21497)

(u'"Amelie (Fabuleux destin d\'Am\xe9lie Poulain', 21434)

(u'Casablanca (1942)', 20939)

(u'Ghostbusters (a.k.a. Ghost Busters) (1984)', 20902)

(u'L\xe9on: The Professional (a.k.a. The Professional) (L\xe9on) (1994)', 20856)

(u'American History X (1998)', 20771)

(u'"Rock', 20339)

(u'Pirates of the Caribbean: The Curse of the Black Pearl (2003)', 20252)

(u'Taxi Driver (1976)', 20210)

(u'Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1964)', 20041)

(u'Babe (1995)', 19916)

(u'"Crouching Tiger', 19626)

(u'"Clockwork Orange', 19608)

(u'Rain Man (1988)', 19570)

(u'"Fifth Element', 19510)

(u'Eternal Sunshine of the Spotless Mind (2004)', 19458)

(u'Trainspotting (1996)', 19337)

(u'Finding Nemo (2003)', 19203)

(u'Apocalypse Now (1979)', 19135)

(u'"Monsters', 19028)

(u'"Shining', 19005)

(u'2001: A Space Odyssey (1968)', 19002)

(u'Mission: Impossible (1996)', 18899)

(u"Ferris Bueller's Day Off (1986)", 18574)

(u'Willy Wonka & the Chocolate Factory (1971)', 18549)

(u'"Truman Show', 18546)

(u"Ocean's Eleven (2001)", 18441)

(u'"Dark Knight', 18345)

(u'X-Men (2000)', 18256)

(u'Full Metal Jacket (1987)', 18232)

(u'Die Hard: With a Vengeance (1995)', 18059)

(u'"Beautiful Mind', 17977)

(u'"Wizard of Oz', 17891)

(u'Shakespeare in Love (1998)', 17816)

(u'Clear and Present Danger (1994)', 17715)

(u'Kill Bill: Vol. 1 (2003)', 17637)

(u'Titanic (1997)', 17631)

(u'Clerks (1994)', 17503)

(u'Mrs. Doubtfire (1993)', 17435)

(u'Four Weddings and a Funeral (1994)', 17417)

(u'"Green Mile', 17404)

(u'Pretty Woman (1990)', 17382)

(u'Minority Report (2002)', 17252)

## Using Broadcast Variables

**#FORMING MOVIE ID AND ITS NAME KEY VALUE PAIR**

moviesfact = moviesrdd.map(lambda x: (x.split(",")[0], x.split(",")[1]))

**#CONVERTING ABOVE KEY VALUE PAIT TO PYTHON’S DICTIONARY:**

dicmovie = moviesfact.collectAsMap()

**#BROADCASTING THE DICTIONARY BY USING SC.BROADCAST () FUNCTION:**

moviebroadcast = sc.broadcast(dicmovie)

>>> moviebroadcast.value["318"]

u'"Shawshank Redemption'

>>> joined = rating.map(lambda (k,v): (k, moviebroadcast.value[k],v)) **#here “rating “ CONTAINS ONLY** **(movieid , count) AND JOINING THE BROADCASTED VARIABLE WITH THE REQUIRED RDD USING .MAP () FUNCTION**

>>> joined.take(5)

[(u'73469', u'Mr. Warmth: The Don Rickles Project (2007)', 20), (u'110008', u'"\'Human\' Factor', 1), (u'101767', u'"Familie', 3), (u'60418', u'Kabluey (2007)', 57), (u'101763', u'My Awkward Sexual Adventure (2012)', 12)]

**#sorting the values**

>>> joinedsorted = joined.sortBy(lambda x:x[2], ascending =False)

>>> joinedsorted.take(10)

[(u'318', u'"Shawshank Redemption', 58207), (u'296', u'Pulp Fiction (1994)', 55266), (u'593', u'"Silence of the Lambs', 54022), (u'356', u'Forrest Gump (1994)', 51707), (u'260', u'Star Wars: Episode IV - A New Hope (1977)', 45572), (u'527', u"Schindler's List (1993)", 44046), (u'2571', u'"Matrix', 44020), (u'110', u'Braveheart (1995)', 42191), (u'50', u'"Usual Suspects', 42105), (u'589', u'Terminator 2: Judgment Day (1991)', 39269)]

**### TOP HUNDERED POPULAR MOVIES**

(u'318', u'"Shawshank Redemption', 58207)

(u'296', u'Pulp Fiction (1994)', 55266)

(u'593', u'"Silence of the Lambs', 54022)

(u'356', u'Forrest Gump (1994)', 51707)

(u'260', u'Star Wars: Episode IV - A New Hope (1977)', 45572)

(u'527', u"Schindler's List (1993)", 44046)

(u'2571', u'"Matrix', 44020)

(u'110', u'Braveheart (1995)', 42191)

(u'50', u'"Usual Suspects', 42105)

(u'589', u'Terminator 2: Judgment Day (1991)', 39269)

(u'457', u'"Fugitive', 38943)

(u'2858', u'American Beauty (1999)', 38399)

(u'1196', u'Star Wars: Episode V - The Empire Strikes Back (1980)', 38350)

(u'1198', u'Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)', 37944)

(u'480', u'Jurassic Park (1993)', 37849)

(u'1', u'Toy Story (1995)', 37494)

(u'858', u'"Godfather', 36742)

(u'1210', u'Star Wars: Episode VI - Return of the Jedi (1983)', 36647)

(u'608', u'Fargo (1996)', 35449)

(u'2959', u'Fight Club (1999)', 35139)

(u'47', u'Seven (a.k.a. Se7en) (1995)', 35056)

(u'150', u'Apollo 13 (1995)', 34705)

(u'32', u'Twelve Monkeys (a.k.a. 12 Monkeys) (1995)', 33414)

(u'2762', u'"Sixth Sense', 32801)

(u'1270', u'Back to the Future (1985)', 31955)

(u'4993', u'"Lord of the Rings: The Fellowship of the Ring', 31935)

(u'2028', u'Saving Private Ryan (1998)', 30454)

(u'5952', u'"Lord of the Rings: The Two Towers', 28839)

(u'590', u'Dances with Wolves (1990)', 28685)

(u'1136', u'Monty Python and the Holy Grail (1975)', 28368)

(u'1197', u'"Princess Bride', 27415)

(u'4226', u'Memento (2000)', 27163)

(u'7153', u'"Lord of the Rings: The Return of the King', 27132)

(u'1193', u"One Flew Over the Cuckoo's Nest (1975)", 26685)

(u'364', u'"Lion King', 26475)

(u'588', u'Aladdin (1992)', 26415)

(u'3578', u'Gladiator (2000)', 26237)

(u'1240', u'"Terminator', 25768)

(u'1291', u'Indiana Jones and the Last Crusade (1989)', 25460)

(u'1214', u'Alien (1979)', 25267)

(u'541', u'Blade Runner (1982)', 25145)

(u'4306', u'Shrek (2001)', 25129)

(u'780', u'Independence Day (a.k.a. ID4) (1996)', 24429)

(u'1265', u'Groundhog Day (1993)', 24244)

(u'1036', u'Die Hard (1988)', 23905)

(u'1221', u'"Godfather: Part II', 23804)

(u'1704', u'Good Will Hunting (1997)', 23499)

(u'1089', u'Reservoir Dogs (1992)', 23468)

(u'380', u'True Lies (1994)', 23447)

(u'377', u'Speed (1994)', 23395)

(u'1213', u'Goodfellas (1990)', 22855)

(u'1580', u'Men in Black (a.k.a. MIB) (1997)', 22692)

(u'1097', u'E.T. the Extra-Terrestrial (1982)', 22550)

(u'1617', u'L.A. Confidential (1997)', 22372)

(u'592', u'Batman (1989)', 22114)

(u'1200', u'Aliens (1986)', 21718)

(u'2997', u'Being John Malkovich (1999)', 21647)

(u'595', u'Beauty and the Beast (1991)', 21497)

(u'4973', u'"Amelie (Fabuleux destin d\'Am\xe9lie Poulain', 21434)

(u'912', u'Casablanca (1942)', 20939)

(u'2716', u'Ghostbusters (a.k.a. Ghost Busters) (1984)', 20902)

(u'293', u'L\xe9on: The Professional (a.k.a. The Professional) (L\xe9on) (1994)', 20856)

(u'2329', u'American History X (1998)', 20771)

(u'733', u'"Rock', 20339)

(u'6539', u'Pirates of the Caribbean: The Curse of the Black Pearl (2003)', 20252)

(u'111', u'Taxi Driver (1976)', 20210)

(u'750', u'Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1964)', 20041)

(u'34', u'Babe (1995)', 19916)

(u'3996', u'"Crouching Tiger', 19626)

(u'1206', u'"Clockwork Orange', 19608)

(u'1961', u'Rain Man (1988)', 19570)

(u'1527', u'"Fifth Element', 19510)

(u'7361', u'Eternal Sunshine of the Spotless Mind (2004)', 19458)

(u'778', u'Trainspotting (1996)', 19337)

(u'6377', u'Finding Nemo (2003)', 19203)

(u'1208', u'Apocalypse Now (1979)', 19135)

(u'4886', u'"Monsters', 19028)

(u'1258', u'"Shining', 19005)

(u'924', u'2001: A Space Odyssey (1968)', 19002)

(u'648', u'Mission: Impossible (1996)', 18899)

(u'2918', u"Ferris Bueller's Day Off (1986)", 18574)

(u'1073', u'Willy Wonka & the Chocolate Factory (1971)', 18549)

(u'1682', u'"Truman Show', 18546)

(u'4963', u"Ocean's Eleven (2001)", 18441)

(u'58559', u'"Dark Knight', 18345)

(u'3793', u'X-Men (2000)', 18256)

(u'1222', u'Full Metal Jacket (1987)', 18232)

(u'165', u'Die Hard: With a Vengeance (1995)', 18059)

(u'4995', u'"Beautiful Mind', 17977)

(u'919', u'"Wizard of Oz', 17891)

(u'2396', u'Shakespeare in Love (1998)', 17816)

(u'349', u'Clear and Present Danger (1994)', 17715)

(u'6874', u'Kill Bill: Vol. 1 (2003)', 17637)

(u'1721', u'Titanic (1997)', 17631)

(u'223', u'Clerks (1994)', 17503)

(u'500', u'Mrs. Doubtfire (1993)', 17435)

(u'357', u'Four Weddings and a Funeral (1994)', 17417)

(u'3147', u'"Green Mile', 17404)

(u'597', u'Pretty Woman (1990)', 17382)

(u'5445', u'Minority Report (2002)', 17252)