~~SE.TER.HIAT.ST~~

SE.TER.HIAT.BA.FE.ZS

SE.TER.HIAT.DO.FE.ZS

SE.TER.HIAT.MS.FE.ZS

SE.TER.CUAT.BA

**For #1 requirement Assumption**

1. The percentages come from the most recent year of that country
2. The graduation rate refers to undergraduates/Bachelors
3. We are excluding post-secondary education statistics because post-secondary includes associate degrees/trade school, and our search is narrow to a minimum of Bachelor’s degree attained

For the 1st business question

We will use CUAT as the primary data, and we’ll disregard the HIAT data for that country

If the country only has HIAT data then we will combine the BA, DO, MA, and ST

CUAT:

Educational attainment, at least Bachelor's or equivalent, population 25+, female (%) (cumulative)

HIAT:

The percentage of population ages 25 and over that attained or completed Bachelor's or equivalent as the highest level of education.

Adam’s PJ2 Q’s:

* where did this data come from?
* How did you handle null’s/NAs?
* What caveats come along with your answer to the above 2 Q’s?

Started PercentageMapper, filter each line to look for either the CUAT statistic or the HIAT statistic, and storing the type of the statistic in a String called “Append”

Then, if value of append is not null (meaning that we are currently on a line that contains our statistic of interest) we iterate through the array we have created of the line of data, and the percentage of the most latest year is stored in a String called “percentage,”

Use context to output the key value pair. The key is the country and the value is the percentage along with the type of percentage it is appended to the end

Hive:

CREATE DATABASE GENDERSTATS\_DB;

USE GENDERSTATS\_DB;

CREATE TABLE GENDER\_STATS (

COUNTRY\_NAME STRING,

COUNTRY\_CODE STRING,

INDICATOR\_NAME1 STRING,

INDICATOR\_NAME2 STRING,

INDICATOR\_NAME3 STRING,

INDICATOR\_NAME4 STRING,

INDICATOR\_CODE STRING,

YEAR\_1960 STRING,

YEAR\_1961 STRING,

YEAR\_1962 STRING,

YEAR\_1963 STRING,

YEAR\_1964 STRING,

YEAR\_1965 STRING,

YEAR\_1966 STRING,

YEAR\_1967 STRING,

YEAR\_1968 STRING,

YEAR\_1969 STRING,

YEAR\_1970 STRING,

YEAR\_1971 STRING,

YEAR\_1972 STRING,

YEAR\_1973 STRING,

YEAR\_1974 STRING,

YEAR\_1975 STRING,

YEAR\_1976 STRING,

YEAR\_1977 STRING,

YEAR\_1978 STRING,

YEAR\_1979 STRING,

YEAR\_1980 STRING,

YEAR\_1981 STRING,

YEAR\_1982 STRING,

YEAR\_1983 STRING,

YEAR\_1984 STRING,

YEAR\_1985 STRING,

YEAR\_1986 STRING,

YEAR\_1987 STRING,

YEAR\_1988 STRING,

YEAR\_1989 STRING,

YEAR\_1990 STRING,

YEAR\_1991 STRING,

YEAR\_1992 STRING,

YEAR\_1993 STRING,

YEAR\_1994 STRING,

YEAR\_1995 STRING,

YEAR\_1996 STRING,

YEAR\_1997 STRING,

YEAR\_1998 STRING,

YEAR\_1999 STRING,

YEAR\_2000 STRING,

YEAR\_2001 STRING,

YEAR\_2002 STRING,

YEAR\_2003 STRING,

YEAR\_2004 STRING,

YEAR\_2005 STRING,

YEAR\_2006 STRING,

YEAR\_2007 STRING,

YEAR\_2008 STRING,

YEAR\_2009 STRING,

YEAR\_2010 STRING,

YEAR\_2011 STRING,

YEAR\_2012 STRING,

YEAR\_2013 STRING,

YEAR\_2014 STRING,

YEAR\_2015 STRING,

YEAR\_2016 STRING)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

TBLPROPERTIES("skip.header.line.count"="1");

DROP TABLE GENDER\_STATS;

CREATE TABLE GENDER\_STATS (

COUNTRY\_NAME STRING,

COUNTRY\_CODE STRING,

INDICATOR\_NAME STRING,

INDICATOR\_CODE STRING,

YEAR\_1960 DECIMAL,

YEAR\_1961 DECIMAL,

YEAR\_1962 DECIMAL,

YEAR\_1963 DECIMAL,

YEAR\_1964 DECIMAL,

YEAR\_1965 DECIMAL,

YEAR\_1966 DECIMAL,

YEAR\_1967 DECIMAL,

YEAR\_1968 DECIMAL,

YEAR\_1969 DECIMAL,

YEAR\_1970 DECIMAL,

YEAR\_1971 DECIMAL,

YEAR\_1972 DECIMAL,

YEAR\_1973 DECIMAL,

YEAR\_1974 DECIMAL,

YEAR\_1975 DECIMAL,

YEAR\_1976 DECIMAL,

YEAR\_1977 DECIMAL,

YEAR\_1978 DECIMAL,

YEAR\_1979 DECIMAL,

YEAR\_1980 DECIMAL,

YEAR\_1981 DECIMAL,

YEAR\_1982 DECIMAL,

YEAR\_1983 DECIMAL,

YEAR\_1984 DECIMAL,

YEAR\_1985 DECIMAL,

YEAR\_1986 DECIMAL,

YEAR\_1987 DECIMAL,

YEAR\_1988 DECIMAL,

YEAR\_1989 DECIMAL,

YEAR\_1990 DECIMAL,

YEAR\_1991 DECIMAL,

YEAR\_1992 DECIMAL,

YEAR\_1993 DECIMAL,

YEAR\_1994 DECIMAL,

YEAR\_1995 DECIMAL,

YEAR\_1996 DECIMAL,

YEAR\_1997 DECIMAL,

YEAR\_1998 DECIMAL,

YEAR\_1999 DECIMAL,

YEAR\_2000 DECIMAL,

YEAR\_2001 DECIMAL,

YEAR\_2002 DECIMAL,

YEAR\_2003 DECIMAL,

YEAR\_2004 DECIMAL,

YEAR\_2005 DECIMAL,

YEAR\_2006 DECIMAL,

YEAR\_2007 DECIMAL,

YEAR\_2008 DECIMAL,

YEAR\_2009 DECIMAL,

YEAR\_2010 DECIMAL,

YEAR\_2011 DECIMAL,

YEAR\_2012 DECIMAL,

YEAR\_2013 DECIMAL,

YEAR\_2014 DECIMAL,

YEAR\_2015 DECIMAL,

YEAR\_2016 DECIMAL)

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

with serdeproperties ( "separatorChar" = ",")

TBLPROPERTIES("skip.header.line.count"="1");

DESCRIBE gender\_stats;

DROP TABLE GENDER\_STATS;

LOAD DATA LOCAL INPATH "/home/cloudera/Downloads/Gender\_StatsData.csv"

INTO TABLE GENDER\_STATS;

SELECT \* FROM GENDER\_STATS;

SELECT \* FROM GENDER\_STATS

WHERE COUNTRY\_NAME = 'United States';

**Query for calculating average yearly percent change in female education across all education levels**

create table US\_Overall\_Female\_Education

AS

select Country\_name, sum(avg\_percent\_change\_yeartoyear)/4 AS US\_yearly\_avg\_percent

FROM US\_Yearly\_Avg\_Percent

group by Country\_name;

**Query for calculating average yearly percent change in female education for primary, secondary, and tertiary education levels**

CREATE TABLE US\_Yearly\_Avg\_Percent\_3

AS

SELECT Country\_Name,Indicator\_Name,Year\_2004,Year\_2015,(((Year\_2015-Year\_2004)/Year\_2004)/11)\*100 AS avg\_percentchange\_yeartoyear

FROM US\_Female\_Education;

**Query for creating a view of just the statistics available for the United States**

CREATE VIEW US\_Female\_Education

(COUNTRY\_NAME,

COUNTRY\_CODE ,

INDICATOR\_NAME ,

INDICATOR\_CODE ,

YEAR\_1960 ,

YEAR\_1961 ,

YEAR\_1962 ,

YEAR\_1963 ,

YEAR\_1964 ,

YEAR\_1965 ,

YEAR\_1966 ,

YEAR\_1967 ,

YEAR\_1968 ,

YEAR\_1969 ,

YEAR\_1970 ,

YEAR\_1971 ,

YEAR\_1972 ,

YEAR\_1973 ,

YEAR\_1974 ,

YEAR\_1975 ,

YEAR\_1976 ,

YEAR\_1977 ,

YEAR\_1978 ,

YEAR\_1979 ,

YEAR\_1980 ,

YEAR\_1981 ,

YEAR\_1982 ,

YEAR\_1983 ,

YEAR\_1984 ,

YEAR\_1985 ,

YEAR\_1986 ,

YEAR\_1987 ,

YEAR\_1988 ,

YEAR\_1989 ,

YEAR\_1990 ,

YEAR\_1991 ,

YEAR\_1992 ,

YEAR\_1993 ,

YEAR\_1994 ,

YEAR\_1995 ,

YEAR\_1996 ,

YEAR\_1997 ,

YEAR\_1998 ,

YEAR\_1999 ,

YEAR\_2000 ,

YEAR\_2001 ,

YEAR\_2002 ,

YEAR\_2003 ,

YEAR\_2004 ,

YEAR\_2005 ,

YEAR\_2006 ,

YEAR\_2007 ,

YEAR\_2008 ,

YEAR\_2009 ,

YEAR\_2010 ,

YEAR\_2011 ,

YEAR\_2012 ,

YEAR\_2013 ,

YEAR\_2014 ,

YEAR\_2015 ,

YEAR\_2016 )

AS

SELECT \* FROM GENDER\_STATS

WHERE COUNTRY\_NAME = 'United States' AND (INDICATOR\_CODE = 'SE.PRM.CUAT.FE.ZS'

OR INDICATOR\_CODE ='SE.SEC.CUAT.LO.FE.ZS' OR INDICATOR\_CODE='SE.SEC.CUAT.UP.FE.ZS'

OR INDICATOR\_CODE='SE.TER.CUAT.ST.FE.ZS');

**Business question #4**

**Indicator Code:** SL.EMP.TOTL.SP.FE.ZS

**Query to isolate the indicator code for the statistic showing female employment (ILO estimate), and also to only portray the statistics for countries and the world (not regions)**

select \* from

(select \*,ROW\_NUMBER() over (order by indicator\_code) as rowid from female\_employment)t where rowid between 0 and 218;

**MySQL**

CREATE TABLE GENDER\_STATS (

COUNTRY\_NAME VARCHAR2(300),

COUNTRY\_CODE STRING,

INDICATOR\_NAME STRING,

INDICATOR\_CODE STRING,

YEAR\_1960 STRING,

YEAR\_1961 STRING,

YEAR\_1962 STRING,

YEAR\_1963 STRING,

YEAR\_1964 STRING,

YEAR\_1965 STRING,

YEAR\_1966 STRING,

YEAR\_1967 STRING,

YEAR\_1968 STRING,

YEAR\_1969 STRING,

YEAR\_1970 STRING,

YEAR\_1971 STRING,

YEAR\_1972 STRING,

YEAR\_1973 STRING,

YEAR\_1974 STRING,

YEAR\_1975 STRING,

YEAR\_1976 STRING,

YEAR\_1977 STRING,

YEAR\_1978 STRING,

YEAR\_1979 STRING,

YEAR\_1980 STRING,

YEAR\_1981 STRING,

YEAR\_1982 STRING,

YEAR\_1983 STRING,

YEAR\_1984 STRING,

YEAR\_1985 STRING,

YEAR\_1986 STRING,

YEAR\_1987 STRING,

YEAR\_1988 STRING,

YEAR\_1989 STRING,

YEAR\_1990 STRING,

YEAR\_1991 STRING,

YEAR\_1992 STRING,

YEAR\_1993 STRING,

YEAR\_1994 STRING,

YEAR\_1995 STRING,

YEAR\_1996 STRING,

YEAR\_1997 STRING,

YEAR\_1998 STRING,

YEAR\_1999 STRING,

YEAR\_2000 STRING,

YEAR\_2001 STRING,

YEAR\_2002 STRING,

YEAR\_2003 STRING,

YEAR\_2004 STRING,

YEAR\_2005 STRING,

YEAR\_2006 STRING,

YEAR\_2007 STRING,

YEAR\_2008 STRING,

YEAR\_2009 STRING,

YEAR\_2010 STRING,

YEAR\_2011 STRING,

YEAR\_2012 STRING,

YEAR\_2013 STRING,

YEAR\_2014 STRING,

YEAR\_2015 STRING,

YEAR\_2016 STRING)

**Q5:**

Adolescent fertility rate (births per 1,000 women ages 15-19)

The percent difference from the year 1960 & 2015

(1960-2015)/1000 \* 100

Indicator code:

SP.ADO.TFRT

1)

CREATE TABLE ADOLESCENT\_FERTILITY\_RATE(

COUNTRY\_NAME VARCHAR(300),

COUNTRY\_CODE VARCHAR(300),

INDICATOR\_NAME VARCHAR(300),

INDICATOR\_CODE VARCHAR(300),

YEAR\_1960 DECIMAL(20,6),

YEAR\_1961 DECIMAL(20,6),

YEAR\_1962 DECIMAL(20,6),

YEAR\_1963 DECIMAL(20,6),

YEAR\_1964 DECIMAL(20,6),

YEAR\_1965 DECIMAL(20,6),

YEAR\_1966 DECIMAL(20,6),

YEAR\_1967 DECIMAL(20,6),

YEAR\_1968 DECIMAL(20,6),

YEAR\_1969 DECIMAL(20,6),

YEAR\_1970 DECIMAL(20,6),

YEAR\_1971 DECIMAL(20,6),

YEAR\_1972 DECIMAL(20,6),

YEAR\_1973 DECIMAL(20,6),

YEAR\_1974 DECIMAL(20,6),

YEAR\_1975 DECIMAL(20,6),

YEAR\_1976 DECIMAL(20,6),

YEAR\_1977 DECIMAL(20,6),

YEAR\_1978 DECIMAL(20,6),

YEAR\_1979 DECIMAL(20,6),

YEAR\_1980 DECIMAL(20,6),

YEAR\_1981 DECIMAL(20,6),

YEAR\_1982 DECIMAL(20,6),

YEAR\_1983 DECIMAL(20,6),

YEAR\_1984 DECIMAL(20,6),

YEAR\_1985 DECIMAL(20,6),

YEAR\_1986 DECIMAL(20,6),

YEAR\_1987 DECIMAL(20,6),

YEAR\_1988 DECIMAL(20,6),

YEAR\_1989 DECIMAL(20,6),

YEAR\_1990 DECIMAL(20,6),

YEAR\_1991 DECIMAL(20,6),

YEAR\_1992 DECIMAL(20,6),

YEAR\_1993 DECIMAL(20,6),

YEAR\_1994 DECIMAL(20,6),

YEAR\_1995 DECIMAL(20,6),

YEAR\_1996 DECIMAL(20,6),

YEAR\_1997 DECIMAL(20,6),

YEAR\_1998 DECIMAL(20,6),

YEAR\_1999 DECIMAL(20,6),

YEAR\_2000 DECIMAL(20,6),

YEAR\_2001 DECIMAL(20,6),

YEAR\_2002 DECIMAL(20,6),

YEAR\_2003 DECIMAL(20,6),

YEAR\_2004 DECIMAL(20,6),

YEAR\_2005 DECIMAL(20,6),

YEAR\_2006 DECIMAL(20,6),

YEAR\_2007 DECIMAL(20,6),

YEAR\_2008 DECIMAL(20,6),

YEAR\_2009 DECIMAL(20,6),

YEAR\_2010 DECIMAL(20,6),

YEAR\_2011 DECIMAL(20,6),

YEAR\_2012 DECIMAL(20,6),

YEAR\_2013 DECIMAL(20,6),

YEAR\_2014 DECIMAL(20,6),

YEAR\_2015 DECIMAL(20,6),

YEAR\_2016 DECIMAL(20,6)

);

2)

sqoop export --connect jdbc:mysql://localhost/GENDERSTATS\_DB --username STUDENT\_ADMIN --password p4ssw0rd --table ADOLESCENT\_FERTILITY\_RATE --input-enclosed-by '"' --fields-terminated-by ',' --lines-terminated-by '\n' --export-dir Project2/Downloads/Gender\_StatsData.csv -m 1

sqoop import --connect jdbc:mysql://localhost/GENDERSTATS\_DB --username STUDENT\_ADMIN --password p4ssw0rd --table ADOLESCENT\_FERTILITY\_RATE --hive-overwrite -m 1

Hive queries:

DROP TABLE if exists fertility\_rate\_stats;

DROP TABLE if exists fertility\_rate\_stats\_3;

CREATE TABLE fertility\_rate\_stats

AS

SELECT \* FROM ADOLESCENT\_FERTILITY\_RATE

WHERE INDICATOR\_CODE='SP.ADO.TFRT';

create table fertility\_rate\_stats\_3

AS

SELECT country\_name, year\_1960, year\_2015, ((year\_1960-year\_2015)/1000)\*100 AS Percent\_Difference

FROM fertility\_rate\_stats;