CA-Drought: Data formatting process 2.0

## A] Urban Water Supplier Report

Update Frequency – Monthly

Current data available till –June 2015

Link – [Download here](http://www.waterboards.ca.gov/water_issues/programs/conservation_portal/conservation_reporting.shtml)

*index="ca\_drought" sourcetype="Water\_GPCD"*

Steps –

1. Download the file from the link above and save it in .csv format. This file will generally have data starting from 2014-July to present.
2. Use the python script UWS formatter.py to format the data in a manner that is consistent with data already indexed in Splunk. This script requires the file name downloaded in step 1 as input and the Month for which the data is to be extracted. The Reporting\_Month should be in the format %y-%b , eg 15-Jun.
3. The formatted file would ideally have these columns. Reporting\_Month and Timestamp should be the first and second columns respectively as there are field extractions defined on it. –

* Reporting\_Month,
* Timestamp
* Supplier\_Name,
* Hydrologic\_Region,
* Total\_Population\_Served,
* Total\_Monthly\_Potable\_Water\_Production\_2014/2015,
* Total\_Monthly\_Potable\_Water\_Production\_2013,
* Units,
* REPORTED\_Residential\_Gallons\_per\_Capita\_Day,
* CALCULATED\_R\_GPCD\_2014/2015,
* CALCULATED\_R\_GPCD\_2013,
* CALCULATED\_Monthly\_Production\_\_Gallons\_2014\_2015,
* CALCULATED\_Monthly\_Production\_\_Gallons\_2013,
* Percent\_Residential\_Use,
* Optional\_Recycled\_Water,
* Recycled\_Water\_Units,
* Comments\_or\_Corrections,
* Stage\_Invoked,
* Mandatory\_Restrictions,
* Optional\_Enforcement\_Actions,
* Optional\_Implementation
* Qualification

**B] Employment data**

Update Frequency – Monthly

Current data available till – April 2015

Link – [Download here](http://www.labormarketinfo.edd.ca.gov/data/unemployment-and-labor-force.html#HIST)

*index="employment\_rate" sourcetype=Employment\_rate*

1. Employment data is available for each county.
2. Copy data starting from the Title column till the last month available and transpose the table while pasting in a new file.
3. Rename the Title field as Year. After the Year column insert two columns titled MSA and MD. If the file has MSA or MD mentioned add it in these columns otherwise insert NULL.
4. Before the Year column insert two columns County and Area. The value for Area column should be same as the one in the original file and the values under County should all end with the “County” eg “Alameda County”.
5. Repeat above process for all 64 files and then run the Format\_Employment\_data.py script which will edit the Industry column names to remove leading & trailing spaces, replace “ “ with “\_”, replace “&” with “and” and other similar transformations. This is done so that its easier to build data models.
6. The script will take as input all the files titled “Input\_data” and store output files in Output\_data.

## C] Demographic data

Update Frequency – Yearly

Current data available till – 2013

Link – [Download here](http://www.census.gov/acs/www/data_documentation/data_profiles/)

*index="demographic" sourcetype=demographic*

Formatting Steps –

1. Rename the entry in A2 as Geo.id and then delete the second column Geo.id2 and delete the first row.
2. Insert the columns from FIPS\_codes.csv after Geo.id :- FIPS\_Code,County,featureId.
3. Before Geo.id add a column titled Year and enter the year in all the rows.
4. Run the python script Format\_Demographic\_data.py, which will remove the columns starting with the word “Percent” and removes all special characters replacing it by “\_”. It takes all the files stored in the folder “Input\_data” as input.
5. Format all the numerical fields and remove the “,” as separator using the format option in Excel.
6. Delete columns BN & BO : Estimate\_RACE\_One\_race and Margin\_of\_Error\_RACE\_One\_race.
7. Delete columns DD & DE : Estimate\_RACE\_Two\_or\_more\_races and Margin\_of\_Error\_RACE\_Two\_or\_more\_races.
8. Rename columns AX, AY,AZ,BA by adding “\_18\_Years\_and\_over”

* Estimate\_SEX\_AND\_AGE\_Male
* Margin\_of\_Error\_SEX\_AND\_AGE\_Male
* Estimate\_SEX\_AND\_AGE\_Female
* Margin\_of\_Error\_SEX\_AND\_AGE\_Female

1. Rename columns BD BE BF BG by adding “\_65\_Years\_and\_over”

* Estimate\_SEX\_AND\_AGE\_Male
* Margin\_of\_Error\_SEX\_AND\_AGE\_Male
* Estimate\_SEX\_AND\_AGE\_Female
* Margin\_of\_Error\_SEX\_AND\_AGE\_Female

1. Delete columns BB, BC, AV,AW –

* Estimate\_SEX\_AND\_AGE\_18\_years\_and\_over
* Margin\_of\_Error\_SEX\_AND\_AGE\_18\_years\_and\_over
* Estimate\_SEX\_AND\_AGE\_65\_years\_and\_over
* Margin\_of\_Error\_SEX\_AND\_AGE\_65\_years\_and\_over

## D] Taxable Sales in California (includes Permits)

Update Frequency – Yearly

Current data available till – 2013

Link – [Download here](http://www.boe.ca.gov/news/tsalescont.htm)

*Tax data - index="taxable\_transactions" sourcetype=taxable\_transactions*

*Permits data - index="permits" sourcetype=permits*

Formatting Steps –

1. The files beginning with “ts\_a<year>.xls” are used as they contain multiple sheets.
2. Start from sheet 3 titled “Table 3. Taxable Sales In The 36 Largest Counties, By Type Of Business”
3. Copy the column all the columns of the first table and paste it into a new sheet.
4. For the next 8 sheets copy only data under the county headers and paste into the sheet created in step 3.
5. Separate (unmerge) the first row and copy paste the county names for the neighboring field ie Taxable Transactions.
6. Format the numbers so that there are no “,” as separators.
7. Add “Total\_” to all the Main categories under Type of Business that are bold formatted.
8. Remove all text formatting and blank rows.
9. Create new sheets to separately copy the columns for Permits and Taxable transactions for each county
10. Once the new sheet is created copy the data and paste its transpose in a new sheet.
11. Before the type of business column (which now has counties). Insert the columns from FIPS\_codes.csv before it :- FIPS\_Code,County,featureId.

Add a column titled Year in the first position. Add a column in the last position titled Tag\_tax/Tag\_permits having the value “Taxable transacations” OR “Permits”. For the column names that show business type replace “ “ with “\_” and remove { . , & ( ) }

1. Repeat the step 10 & 11 for Permits data.
2. Start with **“**Table 4. Taxable Sales In The 22 Smallest Counties, By Type Of Business”
3. Copy the Type of business field for the first table and place data for all the remaining 22 counties adjacent to each other for the two sheets.
4. Add “Total\_” to all the categories under Type of Business. Repeat steps 6 to 12 for the sheet.

**E] Groundwater data**

1. USGS Groundwater data

Update Frequency – Monthly

Current data available till – June 2015

Link – Python script GroundwaterData\_retrieval.py

*index="usgs\_groundwater" sourcetype=csv*

Formatting Steps –

1. Run the python script “GroundwaterData\_retrieval.py” it will require the start and end dates for which data extraction is to be performed.
2. San Bernardino county (FIPS\_CODE = 06071) has a lot of sites so if the time period spans like 6 months or so then for San Bernardino county split it into one month or two months at max.
3. Data will be stored in Output\_data/2015/
4. DWR Groundwater data

Update frequency – Every 3-6 months maybe or whenever they oblige to process the request.

Current data available till – 4th July 2015

Link – Email: [eric.senter@water.ca.gov](mailto:esenter@water.ca.gov)

Ask for data from 4th July onwards.

*index="dwr\_groundwater" sourcetype=DWR\_Groundwater*

The data is well formatted and wont need formatting. Just replace all the spaces in the header columns with ‘\_’.

There is a lookup defined in Splunk which will add the location data to the events.

**F] Water Use**

Update Frequency – Every 5 Years

Current data available till – 2010

Link – [Download here](http://waterdata.usgs.gov/nwis/wu)

*index="water\_use" sourcetype=tsv*

Data has been indexed directly.

**F] Reservoir data –**

Update frequency – Every 3-6 months maybe or whenever they oblige to process the request. Ask for Monthly reservoir storage data as it has data for more reservoirs (around 170+)

Current data available till – June 2015

Link – Email : [welchr@water.ca.gov](mailto:welchr@water.ca.gov)

Ask for data from time period 01/07/2015 to present for the following station IDs

APN,ANT,AST,BRT,BRV,BRD,BTH,BLB,BOC,BMP,BQC,BWN,BWS,BDP,BIO,BUC,BCL,BTV,CCH,CVE,CMN,CFW,ALM,CPL,CSI,CAS,SLW,CHB,CHV,CHY,JNN,CLK,CLA,CGS,CMB,CTG,CYC,COY,MNC,CNV,CUY,MHV,DLV,DMV,DNP,DON,DNN,DNL,DWN,EPK,ELC,ENR,ENG,FRM,FLR,FOL,FRL,FMD,FRD,MIL,GLK,GBR,GBL,PWL,GNT,DAV,HWE,HNS,HID,MEA,HNT,ICH,INP,INL,MMW,INV,IRC,ISB,JCK,ATN,KNT,KRH,KES,LGR,LFY,LGT,HMT,HNN,HDG,LVD,TAH,LVY,LRA,LEW,LNG,LGV,CRW,LON,LSB,LVQ,LWB,CRY,HHL,LYS,SWB,MPL,EDN,MRT,MAT,MHW,MCO,MCS,MMR,MDO,BER,MOR,MRR,NCM,NAT,BUL,EXC,MCR,NHG,NML,NCA,ONF,HTH,ORO,LOT,PAR,HVS,PRR,SCD,PLL,PNF,PT6,PT7,BIT,PVP,PRS,PYM,QUL,RLC,RDN,RLF,RLL,RTD,SDB,SLN,SLS,SNN,SAT,SGB,SNL,SLF,LUS,SPB,SVT,PRU,SGC,SFL,SVO,SHA,SHV,SIV,SKN,SLB,SLC,JNK,SOL,SLJ,SLK,SPG,SPM,STP,STG,SCC,STD,SW3,TRM,TAB,TMT,TNM,CLE,TUL,TLC,TWT,UNV,KLM,SJT,USL,VIL,TAE,EDS,VLP,WRS,WHR,WHI,WSN

Data is well formatted and there is lookup defined to add location information to it also.

OECD

http://www.statista.com/statistics/263156/water-consumption-in-selected-countries/