**Water Quality Dataset:**

The columns in Waterquality.csv are State Name,District Name,Block Name,Panchayat Name,Village Name,Habitation Name,Quality Parameter,Year

1.Remove the headers

2.Copy the local file in hdfs

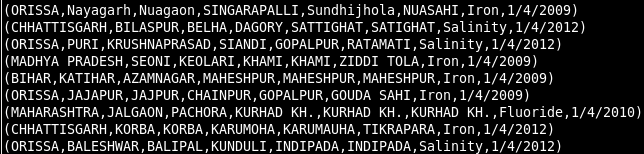
#hadoop dfs -put 'file' /

**LOAD Dataset:**

**** water\_record = LOAD '/Waterquality.csv' using PigStorage(',') as (state:chararray,district:chararray,block:chararray,panchayat:chararray,village:chararray,habit:chararray,quality:chararray,year:chararray);

**View Dataset:**

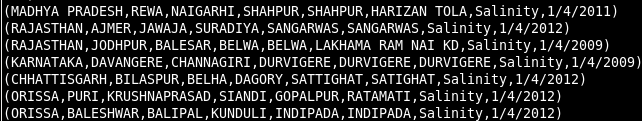
dump water\_record;



**Analysing Dataset:**

**1)Records containing Salinity:**

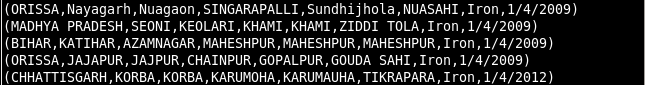
filter\_data = FILTER water\_record by quality=='Salinity';

**** dump filter\_data;

**2)Records Containing Fluoride or Iron:**

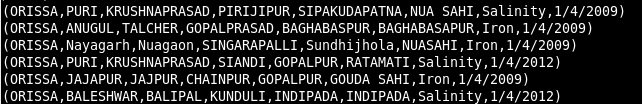
filter\_data = FILTER water\_record by quality in ('Flouride','Iron');

dump filter\_data;



**3)Records containing Salinity or Iron in Orissa:**

filter\_data = FILTER water\_record by quality in ('Salinity','Iron') and state == 'ORISSA';

**** dump filter\_data;

**4)Count of Salinity in each state:**

filter\_data = FILTER water\_record by quality == 'Salinity';

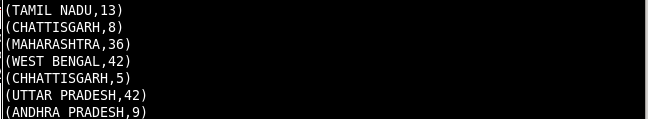
dump filter\_data;

GroupByState = GROUP filter\_data by state;

dump GroupByState;

countbystate = foreach GroupByState GENERATE group,COUNT(filter\_data.state);

dump countbystate;

****

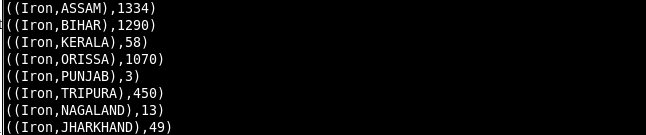
**5)Count of quality in each state:**

**GroupByqualitystate = GROUP water\_record BY (quality,state);**

**dump GroupByqualitystate;**

**countbyqualitystate = foreach GroupByqualitystate GENERATE group,COUNT(water\_record.state);**

**dump countbyqualitystate;**

****