Graded Project

HIVE

Week 8

Table of Content

[Healthcare 3](#__RefHeading___Toc242_2316712879)

[Context 3](#__RefHeading___Toc1022_1253544500)

[Data 3](#__RefHeading___Toc309_2316712879)

[Objective 3](#__RefHeading___Toc311_2316712879)

[Learning Outcomes 3](#__RefHeading___Toc313_2316712879)

[Load Data Into HDFS 3](#__RefHeading___Toc315_2316712879)

[1. Create the table using the above data set, also create the partition based on the time? After creating the partitioned tables, you need to display the partitions. 4](#__RefHeading___Toc244_2316712879)

[2. Total data points collected for each year? 6](#__RefHeading___Toc246_2316712879)

[3. What are different types of cause of death? 6](#__RefHeading___Toc248_2316712879)

[4. Total number of deaths, find out the year in which the deaths are more? 8](#__RefHeading___Toc1084_28677486501)

[5. Top 5 causes of death and Top 5 causes of death across different sex? 9](#__RefHeading___Toc317_2316712879)

# Healthcare

# Context

This data is about the cause of death from 2001 to 2008 in France, this will help the France government to identify the reasons which are the reasons for deaths across different genders and take necessary steps.

# Data

We have a predefined dataset (CauseofDeath.csv) having 4 columns. The dataset has different attributes like.

1. Time(year)

2. SEX (male or female)

3.Cause

4.Value (count of deaths)

# Objective

Show case your hive skills.

# Learning Outcomes

Once you successfully complete this exercise, you will be able to work on dynamic

partitioned tables, windows functions in Hive, using SERDE for building the hive tables.

# [**Load Data Into**](https://www.geeksforgeeks.org/hive-load-data-into-table/) **HDFS**

The first step is to create a folder and upload data into HDFS

**On the CloudX Lab web console:**

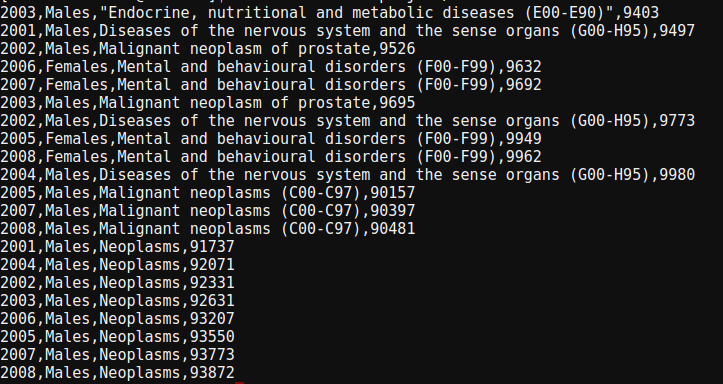
ls

hdfs dfs -ls

hdfs dfs -mkdir project

hdfs dfs -put CauseofDeath.csv project

hdfs dfs -tail project/CauseofDeath.csv



## Create the table using the above data set, also create the partition based on the time? After creating the partitioned tables, you need to display the partitions.

**Create database in hive:**

create database hive\_project;

use hive\_project;

show tables;

**Create table causeOfDeath:**

create table if not exists causeOfDeath (time int, sex STRING, cause

STRING, value int)

row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

with serdeproperties (

'separatorChar'=',',

'quoteChar'='"',

'escapeChar'='\\'

)

stored as textfile

tblproperties('skip.header.line.count'= '1');

**Load the data into the table:**

load data inpath 'project/CauseofDeath.csv' overwrite into table causeOfDeath;

**Create partitioned table:**

create table if not exists cause\_partition(sex STRING, cause

STRING, value int)

partitioned by (time int)

row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

with serdeproperties (

'separatorChar'=',',

'quoteChar'='"',

'escapeChar'='\\'

)

stored as textfile ;

**Configure Hive to allow partitions:**

set hive.exec.dynamic.partition.mode=nonstrict;

set hive.exec.dynamic.partition=true;

**Load data into partition table:**

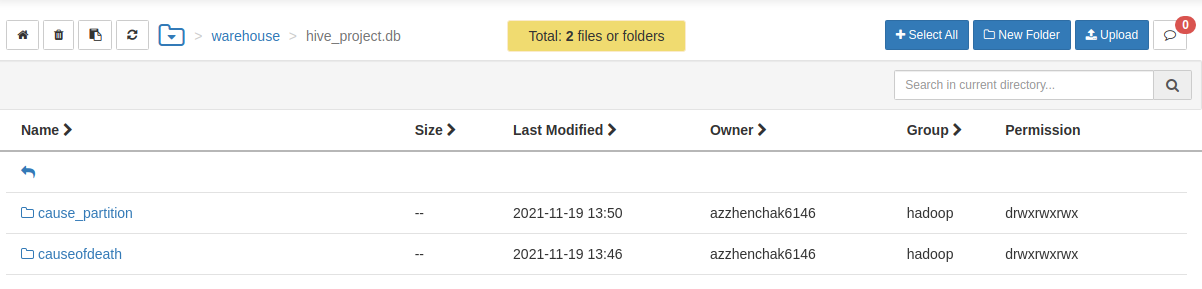
from causeofdeath c INSERT OVERWRITE TABLE cause\_partition

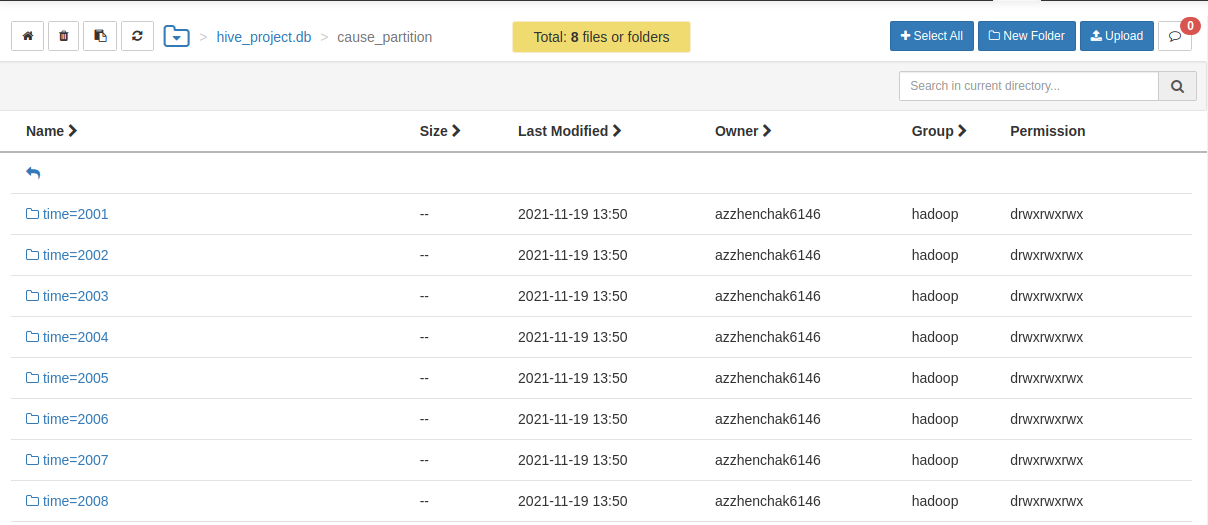
PARTITION(time)

SELECT c.sex,c.cause,c.value,c.time;

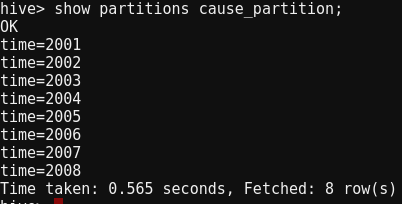
**Check data in Ambari cloudxlab:**

After partition, there were created 8 new files with data corresponding to the year





**Show partitions:**



## Total data points collected for each year?

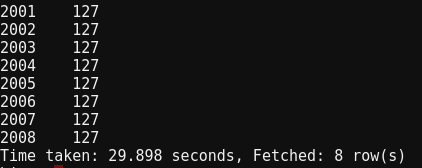
**Counting number data points for each year**

select time, count(\*)

from causeofdeath

group by time;

**Output:**

Data is equally distributed among each year

## What are different types of cause of death?

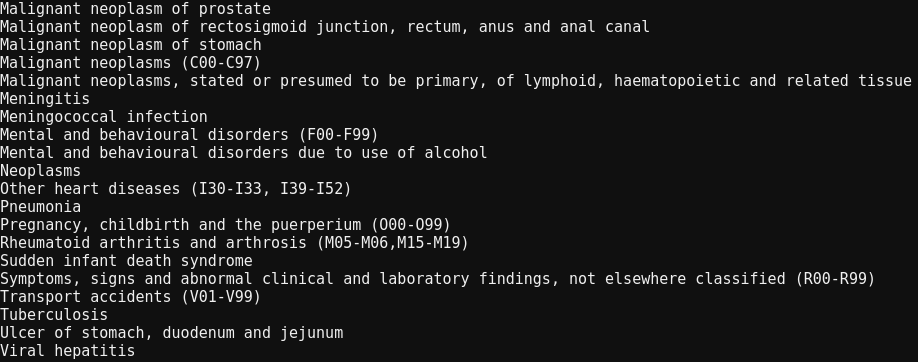
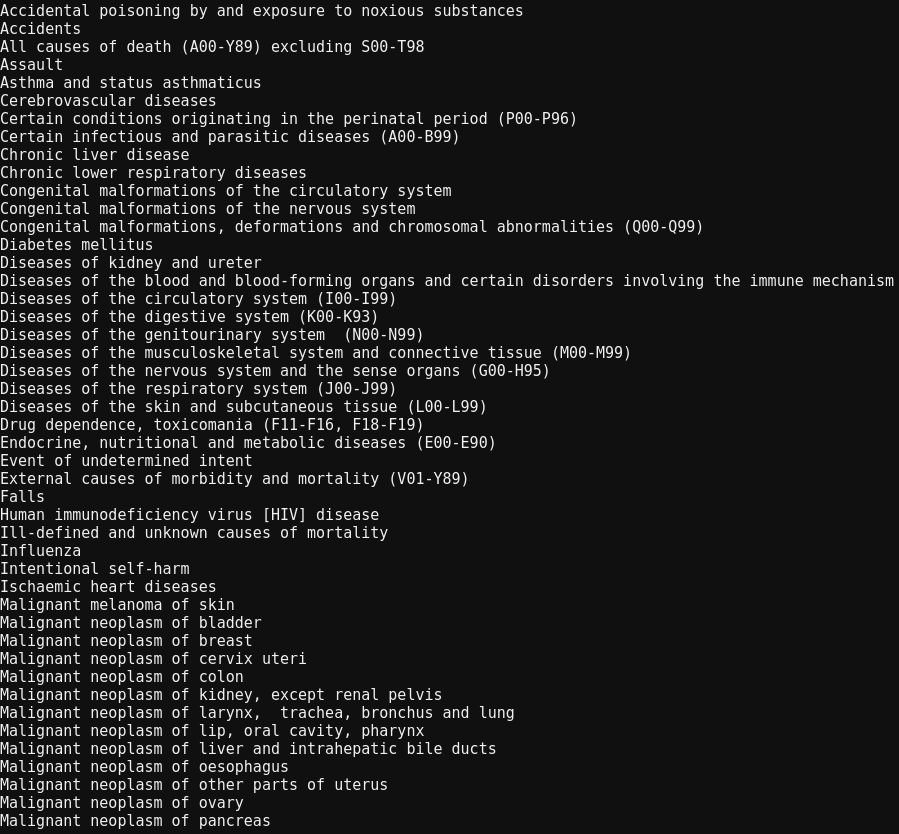
**Types of causes of deaths:**

select cause

from causeofdeath

group by cause;

**Output:**



## Total number of deaths, find out the year in which the deaths are more?

**Total number of deaths:**

select cast(sum(value) as int) from causeofdeath;

**Output:**

12745563

Total number of deaths which happend between 2001 and 2008 is 12745563

**Total number of deaths in each year in descending order:**

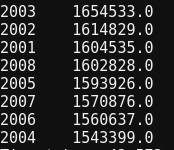
select time, sum(value) as total\_deaths

from causeofdeath

group by time

order by total\_deaths desc;

**Output:**



**Max number of deaths in the year :**

select time, sum(value) as total\_deaths

from causeofdeath

group by time

order by total\_deaths desc

limit 1;

**Output:**

2003 1654533.0

The maximum number of deaths happened during the year 2003 due to various causes.

## Top 5 causes of death and Top 5 causes of death across different sex?

**Top 5 causes of death:**

select sum(value) as sum\_val, cause

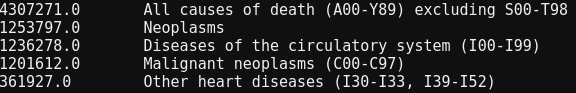
from causeofdeath

group by cause

order by sum\_val desc

limit 5;

**Output:**

“All causes of death (A00-Y89) excluding S00-T98” resulted in 4307271 number of deaths in both genders and stands at the top position for causing most of the deaths, followed by “Neoplasms” which caused 1253797 number of deaths.

**Top 5 causes of death in “Males”:**

select sum(value) as sum\_val, cause

from causeofdeath

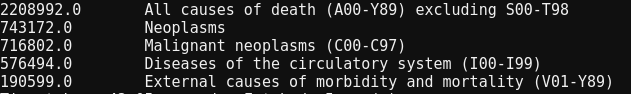
where sex='Males'

group by cause

order by sum\_val desc

limit 5;

**Output:**

“All causes of death (A00-Y89) excluding S00-T98” resulted in 2208992 number of deaths for males and stands at the top position for causing most of the deaths, followed by “Neoplasms” which caused 743172 number of deaths.

**Top 5 causes of death in “Females”:**

select sum(value) as sum\_val, cause

from causeofdeath

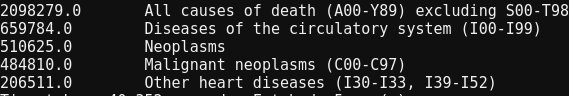
where sex='Females'

group by cause

order by sum\_val desc

limit 5;

**Output:**



“All causes of death (A00-Y89) excluding S00-T98” resulted in 2098279 number of deaths for females and stands at the top position for causing most of the deaths, followed by “Diseases of the circulatory system (I00-I99)” which caused 659784 number of deaths.