

Web Crawling by Using PHP

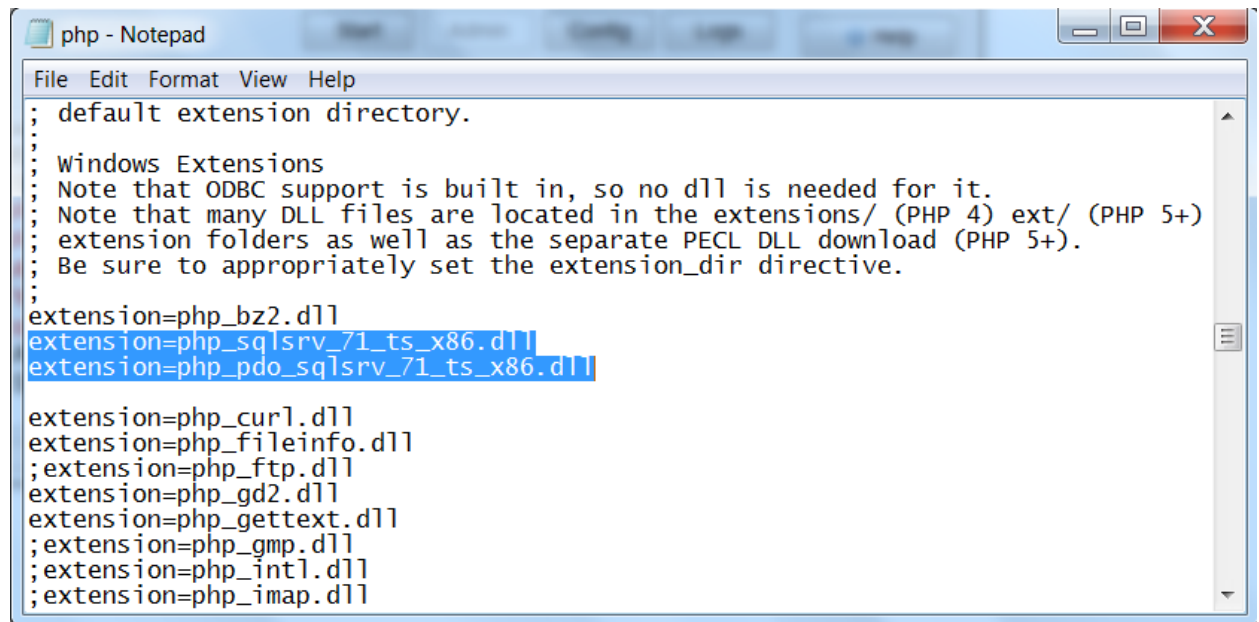
Setting up the drivers

1. Download and install the latest version of web server **XAMPP 7.1.9** which supports **PHP 7.1.9** from <https://www.apachefriends.org/download.html>
2. By default XAMPP runs at **port 80**. Open a web browser and type **localhost:80** then press enter. It should open the XAMPP welcome page. If port 80 is already being used (most probably by Skype) then you can reconfigure XAMPP to port 8080. Once you are able to open the welcome page, click on **phpinfo** to know the PHP version.
3. If the PHP version is 7.1.9 then you need to download MS drivers **SQLSRV43.EXE** for PHP to connect to SQL Server. <https://www.microsoft.com/en-us/download/details.aspx?id=55642>
4. Run SQLSRV43.EXE. When prompted, enter the path to the PHP extensions directory **C:\xampp\php\ext**
5. Open XAMPP control panel. Click the check box on the left of the **Apache** and **SQL** service modules to initialize them. Click on **Config** tab corresponding to Apache and select **php.ini** to open a php file.
6. In the php file find **Windows Extensions** and insert following two drivers as extensions (see the highlighted section in the following figure).

extension=php_sqlsrv_71_ts_x86.dll

extension=php_pdo_sqlsrv_71_ts_x86.dll

Save and close the file.



```
File Edit Format View Help
; default extension directory.
;
; Windows Extensions
; Note that ODBC support is built in, so no dll is needed for it.
; Note that many DLL files are located in the extensions/ (PHP 4) ext/ (PHP 5+)
; extension folders as well as the separate PECL DLL download (PHP 5+).
; Be sure to appropriately set the extension_dir directive.
;
extension=php_bz2.dll
extension=php_sqlsrv_71_ts_x86.dll
extension=php_pdo_sqlsrv_71_ts_x86.dll

extension=php_curl.dll
extension=php_fileinfo.dll
;extension=php_ftp.dll
extension=php_gd2.dll
extension=php_gettext.dll
;extension=php_gmp.dll
;extension=php_intl.dll
;extension=php_imap.dll
```

7. Stop Apache if it is running and then start it again.
8. Open phpinfo (<http://localhost:80/dashboard/phpinfo.php>) and check that both **pdo_sqlsrv** and **sqlsrv** are enabled.

pdo_sqlsrv

pdo_sqlsrv support	enabled
ExtensionVer	4.3.0+9904

sqlsrv

sqlsrv support	enabled
ExtensionVer	4.3.0+9904

Connecting to SQL Server

9. Open **SQL Server Management Studio (SSMS)** and connect the server engine.
10. Copy the following code in Notepad++ and save it as **connect.php** in **htdocs** (C:\xampp\htdocs). All php files need to be saved in htdocs in order to run them in a browser. In the following code use your SQL server name and the correct version of AdventureWorks database installed by you. Note that the purpose of this code is to check the connection with an existing database in SQL server. We will not use AdventureWorks database in this project.

```
<?php
$conn = sqlsrv_connect( "server_name", array( "Database"=>"AdventureWorksDW2014"));
if( $conn ) {
    echo "Connection established.<br />";
}else{
    echo "Connection could not be established.<br />";
    die( print_r( sqlsrv_errors(), true));
}
?>
```

11. Open a web browser and type <http://localhost:80/connect.php> if you are using the default port 80 and press Enter. This will try to connect to the AdventureWorks database installed in the SQL server. If all is well then the browser page will show **Connection established**.

Creating Movie database and a Fact table

12. Click on New Query tab in SSMS. Run the following SQL query to create a database named **movie**.

```
CREATE DATABASE movie;
```

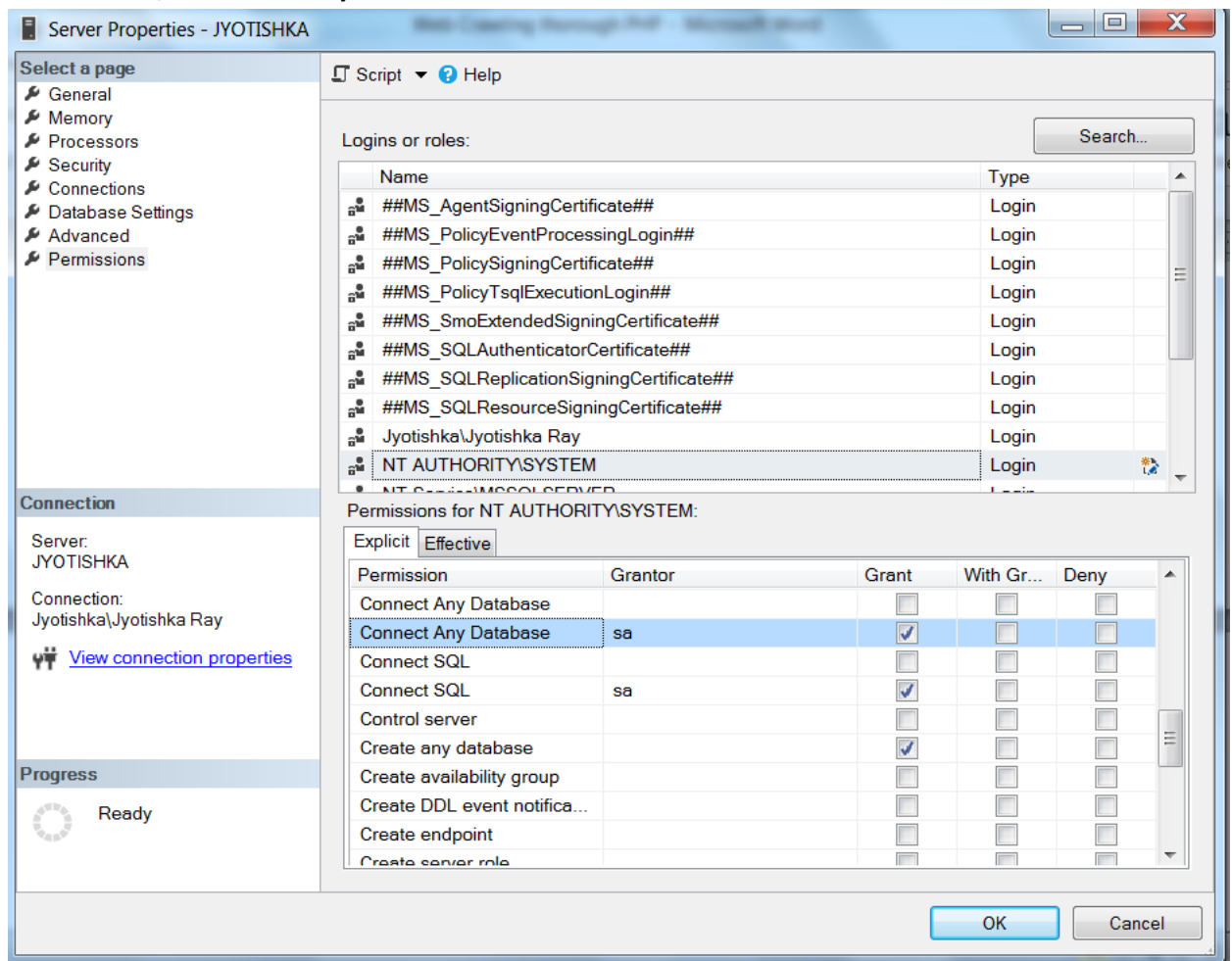
Refresh the **Object Explorer** and observe that a movie database is created.

13. Create a **dbo.Fact** table in movie with following SQL statement.

```
use movie;
create table Fact (ID varchar(255), MovieName varchar(255), Production
varchar(255), Revenue int, No_of_Theaters int, Rev_OpWk int, No_Theaters_OpWk int,
Budget int, MPAA varchar(255), Total_Days int, Genre varchar(255), Closing_Month
int, Closing_Date int, Opening_Month int, Opening_Date int, Release_Year int);
```

Refresh the Object Explorer. Go to the dbo.Fact table in movie database. Expand Columns in Fact table to verify the data type of the columns listed in the Fact table.

- Right-click the movie database and click **Properties** to open **Database Properties**. Click **Permissions** and then click **View server permissions** to open **Server Properties**. Select **NT AUTHORITY\SYSTEM**. In the **Explicit** tab check the **Grant** boxes for **Connect Any Database**, **Connect SQL**, and **Create any database** as shown below. Click OK.



- Right-click the **dbo.Fact** table and click **Properties** to open **Table Properties**. Click **Permissions** and then click **Browse**. Select **[public]** by checking the box. Click OK twice. With public selected in the **Users or roles** check the **Grant** boxes for **Alter**, **Control**, **Delete**, **Insert**, **Select**, **Take ownership**, and **Update**. Do the same by clicking **View schema permissions**. Click OK.

Add Data

- Open the following movie website.

<http://www.boxofficemojo.com/yearly/chart/?page=1&view=releasedate&view2=domestic&yr=2010&p=.htm>

The website contains 100 movies in each webpage with top domestic grosses. You can click any movie link to open a new webpage with more details about the movie. We will collect the information of top 100 movies in years 2010 to 2014 and save it in the movie database.

17. Create a folder named **Project 1** in **htdocs**. Inside Project 1, create an empty folder named **ID**. Modify the **movie.php** file provided to you by changing the **Server_Name** and save it in **htdocs**. When the connection is established, run the **movie.php** file in **localhost**. This will add the data of 100 movies with highest domestic grosses in years 2010 to 2014. Note that the **Apache** must be running in XAMPP and a connection must be established with the SQL server to save the data in the server.
18. Run the SQL query `select * from Fact;` to view the data in SSMS.