Session 10

Assignment 2

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 10 – Assignment 2 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 10 – Assignment 2** |
|  |  |
| **Start Date:** | 05/10/2017 |
|  |  |
|  |  |



# 

# Contents

[Contents 2](#_Toc494934088)

[Change History 3](#_Toc494934089)

[1. Problem Statement 4](#_Toc494934090)

[2. Loading Data Into HBase Using PIG Scripts. 5](#_Toc494934091)

[2.1. Copying the data set in to HDFS 5](#_Toc494934092)

[2.2. Adding Jars 5](#_Toc494934093)

[2.3. Starting HBASE 6](#_Toc494934094)

[2.4. Pig 6](#_Toc494934095)

[2.5. Check results 7](#_Toc494934096)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 05/10/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

Implement the concept given in below blog link and share the complete steps along with

screenshots.

*https://acadgild.com/blog/loading-data-into-hbase-using-pig-scripts/*

# Loading Data Into HBase Using PIG Scripts.

In this assignment we will be discussing the loading of data into HBase using Pig scripts.

To implement the concepts an Hadoop cluster with Pig and HBase running on it was used.

This assignment will demonstrate step by step clarification regarding transferring data into HBase using Pig.

We are taking sample data set of student which will be loaded into HBase.

**Sample of Dataset**

StudentName,sector,DOB,qalification,score,state,randomName

ABROSER,goverenment,18-11-2002,MBBS,3.5,Pennsylvania,prattville\*

ALEXANDER,goverenment,20-10-2000,BSC,2.5,vermont,gadsden+

ALEXANDER,private,20-10-2000,BE,8.5,arizona,decatur!

ALEXANDER,goverenment,01-01-2003,BTECH,4.5,oregon,huntsville/

AGNEW,goverenment,20-10-2000,BCOM,7.5,california,dothan@

ATNEST,goverenment,20-10-2000,MTECH,8.5,arizona,decatur!

BELL,goverenment,10-07-2004,BBA,9.5,alaska,auburn~

BURR,goverenment,12-12-2001,BE,100,alabama,madison`

BURD,goverenment,20-10-2000,ME,6.5,louisiana,hoover#

BACHTEL,goverenment,28-04-2005,BE,100,alabama,madison`

Please refer the description for the above data set containing  seven columns named as:

**StudentName, sector, DOB, qualification, score, state, randomName.**

## Copying the data set in to HDFS

The dataset was copied into HDFS

*[acadgild@localhost ~]$ hadoop fs -copyFromLocal student.txt /user/acadgild/Dataset/student.txt*

## Adding Jars

We will be including few jar files of HBase to the Pig classpath.

*[acadgild@localhost ~]$ PIG\_CLASSPATH=/usr/local/hbase/lib*

## Starting HBASE

We will now start HBase shell and create a table.

We only need this table as skeleton so PIG can Store data inside this by referring the table name.

*hbase(main):007:0>*

*hbase(main):007:0> create 'studentAcad1','student\_data'*

*0 row(s) in 0.5520 seconds*

*=> Hbase::Table - studentAcad1*

*hbase(main):008:0>*

Come out from HBase by typing exit and switch to PIG grunt shell.

## Pig

Once we are inside PIG mode we can load data from HDFS to Alias relation.

Now we can transfer the data inside HBase by STORE command.

We need to ensure that we give the correct name for table name created inside HBase. Also the parameters should be kept in mind to avoid mistake.

*rData =LOAD '/user/acadgild/Dataset/student.txt' USING PigStorage(',') AS (StudentName:chararray,sector:chararray,DOB:chararray,qualification:chararray,score:int,state:chararray,randomName:chararray);*

*Store rData INTO 'hbase:studentAcad' USING org.apache.pig.backend.hadoop.hbase.HBaseStorage('student\_data:StudentName,student\_data:sector,student\_data:DOB,student\_data:qualification,student\_data:score,student\_data:state,student\_data:randomName');*

## Check results

**Checking database**

The result can be displayed through scan command followed by table name inside quotes( ‘ ‘ ).

*hbase(main):006:0> scan 'studentAcad'*

*ROW COLUMN+CELL*

*ABEDNIGO column=student\_data:DOB, timestamp=1506580157699, value=BB*

*A*

*ABEDNIGO column=student\_data:StudentName, timestamp=1506580157699,*

*value=goverenment*

*ABEDNIGO column=student\_data:qualification, timestamp=1506580157699*

*, value=100*

*ABEDNIGO column=student\_data:score, timestamp=1506580157699, value=*

*alabama*

*ABEDNIGO column=student\_data:sector, timestamp=1506580157699, value*

*=20-10-2000*

*ABEDNIGO column=student\_data:state, timestamp=1506580157699, value=*

*madison`*

*ABROSER column=student\_data:DOB, timestamp=1506580157622, value=MB*

*BS*

*ABROSER column=student\_data:StudentName, timestamp=1506580157622,*

*value=goverenment*

*ABROSER column=student\_data:qualification, timestamp=1506580157622*

*, value=3*

*ABROSER column=student\_data:score, timestamp=1506580157622, value=*

*Pennsylvania*

*…...................*

*230 row(s) in 0.6380 seconds*

**Checking 1 Record**

*hbase(main):006:0> scan 'studentAcad'get 'studentAcad','ABROSER'*

*COLUMN CELL*

*student\_data:DOB timestamp=1506580157622, value=MBBS*

*student\_data:StudentName timestamp=1506580157622, value=goverenment*

*student\_data:qualification timestamp=1506580157622, value=3*

*student\_data:score timestamp=1506580157622, value=Pennsylvania*

*student\_data:sector timestamp=1506580157622, value=18-11-2002*

*student\_data:state timestamp=1506580157622, value=prattville\**

*6 row(s) in 0.0220 seconds*