**Excercise one:**

Create a Directory in HDFS using Ambari e.g "user/maria\_dev/pig\_demo"

The input file of Pig contains each tuple/record in individual lines. And the entities of the record are separated by a delimiter (In our example we used **“,”**).

In the local file system, create an input file **student\_data.txt** containing data as shown below.

001,Rajiv,Reddy,9848022337,Hyderabad

002,siddarth,Battacharya,9848022338,Kolkata

003,Rajesh,Khanna,9848022339,Delhi

004,Preethi,Agarwal,9848022330,Pune

005,Trupthi,Mohanthy,9848022336,Bhuwaneshwar

006,Archana,Mishra,9848022335,Chennai.

**Execute the Load Statement**

Now load the data from the file **student\_data.txt** into Pig by executing the following Pig

Latin statement

1. student = LOAD '/user/maria\_dev/pig\_demo/student\_data.txt' USING PigStorage(',') as ( id:int, firstname:chararray, lastname:chararray, phone:chararray, city:chararray );
2. Dump student;

Following is the description of the above statement.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| column | id | firstname | lastname | phone | city |
| datatype | int | char array | char array | char array | char array |

|  |  |
| --- | --- |
| Relation name | We have stored the data in the schema **student**. |
| Input file path | We are reading data from the file **student\_data.txt**, which is in the /pig\_data/ directory of HDFS. |
| Storage function | We have used the **PigStorage()** function. It loads and stores data as structured text files. It takes a delimiter using which each entity of a tuple is separated, as a parameter. By default, it takes ‘\t’ as a parameter. |
| schema | We have stored the data using the following schema. |

**Note:** The **load** statement will simply load the data into the specified relation in Pig

Storing Data

You can store the loaded data in the file system using the **store** operator.

**Syntax**

Given below is the syntax of the Store statement.

STORE Relation\_name INTO ' required\_directory\_path ' [USING function];

Now, let us store the relation in the HDFS directory

STORE student INTO '/user/maria\_dev/pig\_outout' USING PigStorage (',');

**Group Key word Exercise**

1. student\_details = LOAD '/user/maria\_dev/student\_data.txt' USING PigStorage(',') as (id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray, city:chararray);
2. group\_data = GROUP student\_details by age;
3. Dump group\_data;

Grouping Two Relations using Cogroup

Assume that we have two files namely **student\_details.txt** and **employee\_details.txt** in the HDFS directory **/pig\_data/** as shown below.

**student\_details.txt**

001,Rajiv,Reddy,21,9848022337,Hyderabad

002,siddarth,Battacharya,22,9848022338,Kolkata

003,Rajesh,Khanna,22,9848022339,Delhi

004,Preethi,Agarwal,21,9848022330,Pune

005,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar

006,Archana,Mishra,23,9848022335,Chennai

007,Komal,Nayak,24,9848022334,trivendram

008,Bharathi,Nambiayar,24,9848022333,Chennai

**employee\_details.txt**

001,Robin,22,newyork

002,BOB,23,Kolkata

003,Maya,23,Tokyo

004,Sara,25,London

005,David,23,Bhuwaneshwar

006,Maggy,22,Chennai

And we have loaded these files into Pig with the relation names **student\_details** and **employee\_details** respectively, as shown below.

student\_details = LOAD '/user/maria\_dev/pig\_data/student\_details.txt' USING PigStorage(',')

as (id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray, city:chararray);

employee\_details = LOAD '/user/maria\_dev/pig\_data/employee\_details.txt' USING PigStorage(',')

as (id:int, name:chararray, age:int, city:chararray);

Now, let us group the records/tuples of the relations **student\_details** and **employee\_details** with the key age, as shown below.

cogroup\_data = COGROUP student\_details by age, employee\_details by age;

Verification

Verify the relation **cogroup\_data** using the **DUMP** operator as shown below.

Dump cogroup\_data;

**Join Exercise**

1. student = LOAD '/user/maria\_dev/student\_data.txt' USING PigStorage(',') as (id:int,name:chararray,city:chararray);
2. student2 = LOAD '/user/maria\_dev/student\_data2.txt' USING PigStorage(',') as (id:int,name:chararray,city:chararray);
3. e = join student by id, student2 by id;
4. DESCRIBE e;
5. Dump e;