Hadoop: Hands-On:
Pig Commands

Pig Commands:

Command 1: Start Pig in the local mode

```
$pig -x local
```

Command 2: Start Pig in the MapReduce mode

```
$pig -x mapreduce
```

Command 3: Load the input data (You can replace 'Student_info' with another file)

Content for the dataset (Copy and Paste within a Student_info.txt file):

```
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.
```

```
GRUNT> A = LOAD 'Student_info' using PigStorage(',');
 GRUNT> B = LIMIT log 4; // limits the record to 4 rows
```

```
grunt> A = load 'Student_info' using PigStorage(',');
2019-11-27 06:15:43,570 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = limit A 4;
```



```
(001,Rajiv,Reddy,9848022337,Hyderabad)
(002,siddarth,Battacharya,9848022338,Kolkata)
(003,Rajesh,Khanna,9848022339,Delhi)
(004,Preethi,Agarwal,9848022330,Pune)
```

Command 4: Display your data

```
GRUNT>dump A;
```

```
grunt> A = load 'Student_info' using PigStorage(':');
2019-11-27 06:55:17,723 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-11-27 06:55:17,724 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> dump A;
```

Command 5: Store the data

```
grunt> STORE student INTO '<your storage directory>' USING PigStorage
 (',');
```

Command 6: Filter the data (an example)

```
grunt>A = load 'Student_info' as (Roll:int, name:chararray,
  lastname:chararray, number:int ,city:chararray);
grunt>B = FILTER 'Student_info' BY city == 'Kolkata';
  grunt> dump B;
```

```
grunt> A = load 'Student_info' as (Roll:int, name:chararray, lastname:chararray,
 number:int ,city:chararray);
2019-11-27 06:47:27,735 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-11-27 06:47:27,735 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = FILTER A BY city == 'Kolkata';
grunt> dump B;
```

```
(006,Archana,Mishra,9848022335,Kolkata)
(002,siddarth,Battacharya,9848022338,Kolkata)
grunt>
```


Command 7: Group the data (an example)

```
grunt>grouped_records = GROUP college_students by number; grunt>
 dump grouped_records;
```

```
grunt> A = load 'Student_info' using PigStorage (',') AS (Roll:int, name:chararray, la
stname:chararray, number:int ,city:chararray);
2019-11-27 07:59:17,539 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-11-27 07:59:17,539 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> grouped_records = GROUP A by city;
grunt> dump grouped_records;
```

```
(Pune,{(4,Preethi,Agarwal,,Pune)})
(Delhi,{(3,Rajesh,Khanna,,Delhi)})
(Kolkata,{(2,siddarth,Battacharya,,Kolkata)})
(Kolkata.,{(6,Archana,Mishra,,Kolkata.)})
(Hyderabad,{(1,Rajiv,Reddy,,Hyderabad)})
(Bhuwaneshwar,{(5,Trupthi,Mohanthy,,Bhuwaneshwar)})
grunt>
```

Command 8: Remove the duplicate tuples

```
grunt> alias=DISTINCT alias;
 grunt> dump A;
```

```
grunt> A = load 'Student_info' using PigStorage(',');
2019-12-23 05:17:43,485 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-12-23 05:17:43,486 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = DISTINCT A;
grunt> Dump B;
```

```
(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,Kolkata)
grunt>
```


Command 9: Perform a Join

```
grunt>JOIN alias BY {expression|'('expression [, expression …]')'} (, alias
  BY {expression|'('expression [, expression …]')'} …)
```

```
r the session: PIG-default-432cc53b-d2c8-419c-ac7c-0d0ad41234f5
2019-12-01 23:56:44,205 [main] WARN  org.apache.pig.PigServer - ATS is disabled
since yarn.timeline-service.enabled set to false
grunt> customer = LOAD 'customers.txt' USING PigStorage(',') as (id:int, name:ch
ararray, age:int, address:chararray, salary:int);
2019-12-01 23:57:25,199 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
```

```
2019-12-01 23:57:25,199 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> orders = LOAD 'orders.txt' USING PigStorage(',') as (id:int,  date:charar
ray, customer id:int, amount:int);
```

```
Details at logfile: /home/hadoop/pig_1575262603002.log
grunt> customer_orders = JOIN customer BY id, orders BY id;
grunt> DUMP customer_orders;
```

Command 10: Co-group the data

```
grunt>cogroup_data= COGROUPstudent_details by age, employee_details
  by age;
```

```
grunt> student_details = LOAD 'Student_details' USING PigStora
ge(',') as (id:int, firstname:chararray, lastname:chararray, a
ge:int, phone:chararray, city:chararray);
2019-12-23 05:45:47,060 [main] INFO  org.apache.hadoop.conf.Co
nfiguration.deprecation - io.bytes.per.checksum is deprecated.
 Instead, use dfs.bytes-per-checksum
2019-12-23 05:45:47,061 [main] INFO  org.apache.hadoop.conf.Co
nfiguration.deprecation - fs.default.name is deprecated. Inste
ad, use fs.defaultFS
grunt> employee_details = LOAD 'employee_details' USING PigSto
rage(',') as (id:int, name:chararray, age:int, city:chararray)
;
2019-12-23 05:46:00,353 [main] INFO  org.apache.hadoop.conf.Co
nfiguration.deprecation - io.bytes.per.checksum is deprecated.
 Instead, use dfs.bytes-per-checksum
2019-12-23 05:46:00,353 [main] INFO  org.apache.hadoop.conf.Co
nfiguration.deprecation - fs.default.name is deprecated. Inste
ad, use fs.defaultFS
grunt>
```

```
grunt> cogroup_data = COGROUP student_details by age, employee_details by age;
grunt> dump cogroup_data;
```



```
(21,{(6,Archana,Mishra,21,9848022335,Chennai),(4,Reenu,Agarwal,21,9848022330,Pune)
},{(3,Maya,21,Tokyo )})
(22,{(3,Rajesh,Khanna,22,9848022339,Delhi),(2,Siddarth,Battacharya,22,9848022338,K
olkata)},{(6,Maggy,22,Chennai),(1,Robin,22,newyork )})
(23,{(5,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar),(1,Yash,Reddy,23,9848022337,H
yderabad)},{(5,David,23,Bhuwaneshwar ),(2,BOB,23,Kolkata )})
(24,{(8,Bharathi,Nambiayar,24,9848022333,Chennai),(7,Komal,Nayak,24,9848022334,tri
vendram)},{})
(25,{},{(4,Sara,25,London )})
grunt>
```

Command 11: Compute the cross product

```
grunt> Relation3_name = CROSS Relation1_name, Relation2_name;
```

```
grunt> orders = LOAD 'order' USING PigStorage(',') as (oid:int, date:chararray,
customer_id:int, amount:int);
2019-12-23 06:46:55,300 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-12-23 06:46:55,300 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> customers = LOAD 'customer.txt' USING PigStorage(',') as (id:int, name:ch
ararray, age:int, address:chararray, salary:int);
2019-12-23 06:47:15,084 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-12-23 06:47:15,085 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> cross_data = CROSS customers, orders;
grunt> dump cross_data;
```

```
(7,Muffy,24,Indore,10000,103,2018-05-20 00:00:00,4,2060)
(7,Muffy,24,Indore,10000,101,2019-11-20 00:00:00,2,1560)
(7,Muffy,24,Indore,10000,100,2019-10-08 00:00:00,3,1500)
(7,Muffy,24,Indore,10000,102,2019-10-08 00:00:00,3,3000)
(6,Komal,22,MP,4500,103,2018-05-20 00:00:00,4,2060)
(6,Komal,22,MP,4500,101,2019-11-20 00:00:00,2,1560)
(6,Komal,22,MP,4500,100,2019-10-08 00:00:00,3,1500)
(6,Komal,22,MP,4500,102,2019-10-08 00:00:00,3,3000)
(5,Hardik,27,Bhopal,8500,103,2018-05-20 00:00:00,4,2060)
(5,Hardik,27,Bhopal,8500,101,2019-11-20 00:00:00,2,1560)
(5,Hardik,27,Bhopal,8500,100,2019-10-08 00:00:00,3,1500)
(5,Hardik,27,Bhopal,8500,102,2019-10-08 00:00:00,3,3000)
(4,Chaitali,25,Mumbai,6500,103,2018-05-20 00:00:00,4,2060)
(4,Chaitali,25,Mumbai,6500,101,2019-11-20 00:00:00,2,1560)
(4,Chaitali,25,Mumbai,6500,100,2019-10-08 00:00:00,3,1500)
(4,Chaitali,25,Mumbai,6500,102,2019-10-08 00:00:00,3,3000)
(3,Kaushik,23,Kota,2000,103,2018-05-20 00:00:00,4,2060)
(3,Kaushik,23,Kota,2000,101,2019-11-20 00:00:00,2,1560)
(3,Kaushik,23,Kota,2000,100,2019-10-08 00:00:00,3,1500)
(3,Kaushik,23,Kota,2000,102,2019-10-08 00:00:00,3,3000)
(2,Khilan,34,Chennai,11500,103,2018-05-20 00:00:00,4,2060)
(2,Khilan,34,Chennai,11500,101,2019-11-20 00:00:00,2,1560)
(2,Khilan,34,Chennai,11500,100,2019-10-08 00:00:00,3,1500)
(2,Khilan,34,Chennai,11500,102,2019-10-08 00:00:00,3,3000)
(1,Ram,43,Delhi,12000,103,2018-05-20 00:00:00,4,2060)
(1,Ram,43,Delhi,12000,101,2019-11-20 00:00:00,2,1560)
(1,Ram,43,Delhi,12000,100,2019-10-08 00:00:00,3,1500)
(1,Ram,43,Delhi,12000,102,2019-10-08 00:00:00,3,3000)
grunt>
```

