**ACD\_BDD\_Session\_8\_Assignment\_2\_Main**

**Problem Statement**

**Create a sample dataset and implement the below Pig commands on the same dataset.**

1) Concat

2) Tokenize

3) Sum

4) Min

5) Max

6) Limit

7) Store

8) Distinct

9) Flatten

10) IsEmpty

**Solution**

Data stored on local: /usr/local/hadoop-2.6.0/sbin/functions.txt'

Bob Goldfish 5  
Alice Dog 3  
Bob Cat 4  
Alice Cat 2  
Tom mice 3  
 Dog 6

grunt> data = load '/usr/local/hadoop-2.6.0/sbin/functions.txt' AS (f1:chararray, f2:chararray, f3:chararray);  
  
grunt> DUMP data;  
  
  
(Alice,Turtle,1)  
(Bob,Goldfish,5)  
(Alice,Dog,3)  
(Bob,Cat,4)  
(Alice,Cat,2)  
(Tom,mice,3)  
(,Dog,6)

grunt> **STORE** data INTO 'storedata';

grunt> DESCRIBE data;  
  
data: {f1: chararray,f2: chararray,f3: int}

grunt> C = FOREACH data GENERATE **CONCAT**(f1,f2);

grunt> DUMP C;

(AliceTurtle)  
(BobGoldfish)  
(AliceDog)  
(BobCat)  
(AliceCat)  
(Tommice)  
(Dog)  
  
grunt> words = FOREACH data GENERATE **flatten**(**TOKENIZE**(f1)) AS word;  
  
grunt> DUMP words;  
  
grunt> G = GROUP data BY f1;  
  
grunt> DUMP G;  
  
(Bob,{(Bob,Cat,4),(Bob,Goldfish,5)})  
(Tom,{(Tom,mice,3)})  
(Alice,{(Alice,Cat,2),(Alice,Dog,3),(Alice,Turtle,1)})  
(,{(,Dog,6)})  
  
grunt> DESCRIBE G;  
G: {group: chararray,data: {(f1: chararray,f2: chararray,f3: int)}}  
  
grunt> S = FOREACH G GENERATE group, **SUM**(data.f3);  
  
(Bob,9)  
(Tom,3)  
(Alice,6)  
(,6)  
  
grunt> DESCRIBE S;  
S: {group: chararray,long}  
  
grunt> Max = FOREACH G GENERATE group, **MAX**(data.f3);  
  
grunt> DUMP Max;  
  
(Bob,5)  
(Tom,3)  
(Alice,3)  
(,6)  
  
  
grunt> Min = FOREACH G GENERATE group, **MIN**(data.f3);  
  
grunt> DUMP Min;  
  
(Bob,4)  
(Tom,3)  
(Alice,1)  
(,6)  
  
  
grunt> D = **DISTINCT** data;  
  
(Bob,Cat,4)  
(Bob,Goldfish,5)  
(Tom,mice,3)  
(Alice,Cat,2)  
(Alice,Dog,3)  
(Alice,Turtle,1)  
(,Dog,6)  
  
  
grunt> L = **LIMIT** data 3;  
  
(Alice,Turtle,1)  
(Bob,Goldfish,5)  
(Alice,Dog,3)