**1. Introduction**

This assignment will help you to consolidate the concepts learnt in the session.

**2. 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

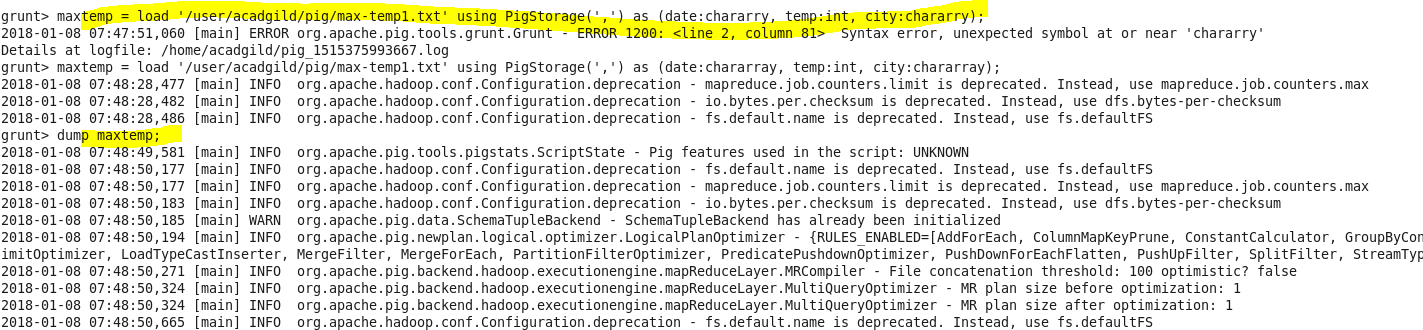
8) Distinct

9) Flatten

10) IsEmpty

Reach the grant shell by using PIG

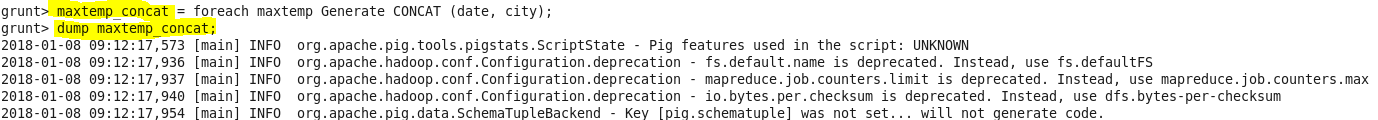
loading the dataset from max-temp.txt

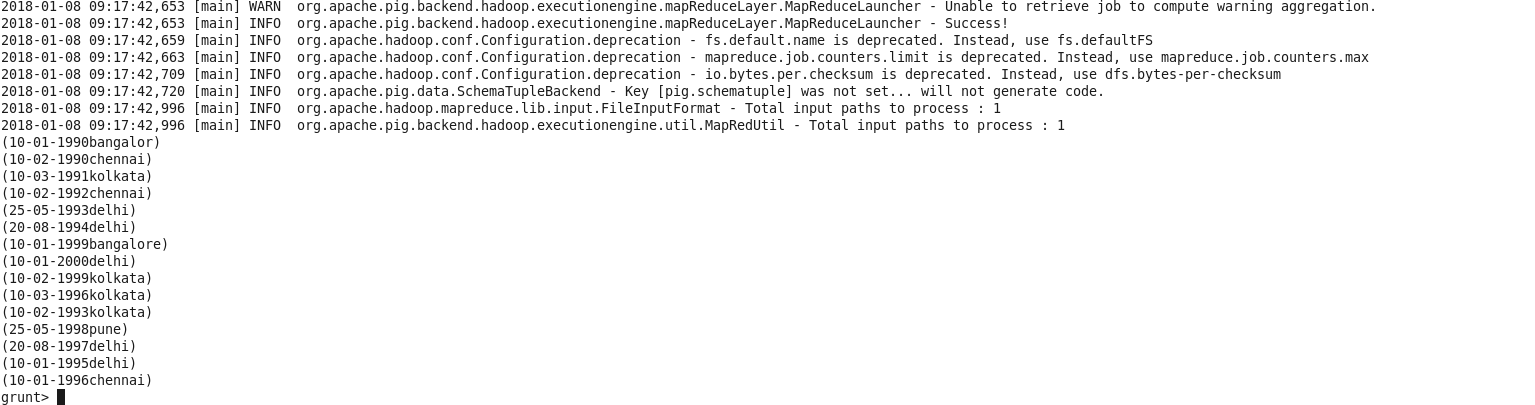




CONCAT:

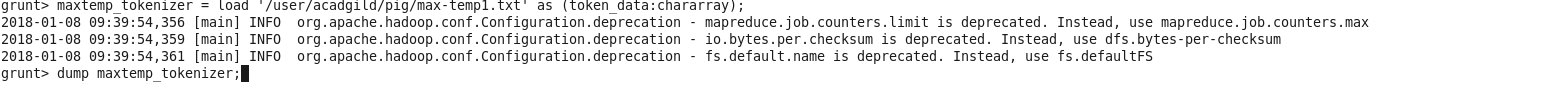
Concat is use for concatenation .





**ToKenize**

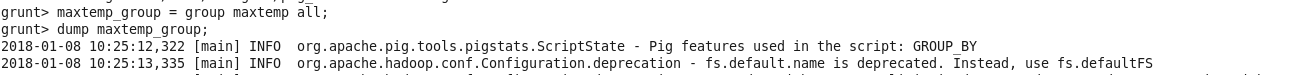
The **TOKENIZE()** function of Pig Latin is used to split a string (which contains a group of words) in a single tuple and returns a bag which contains the output of the split operation.



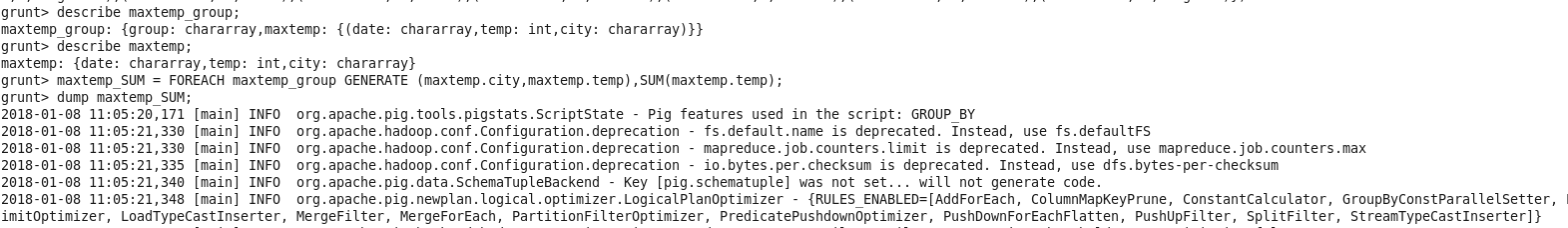


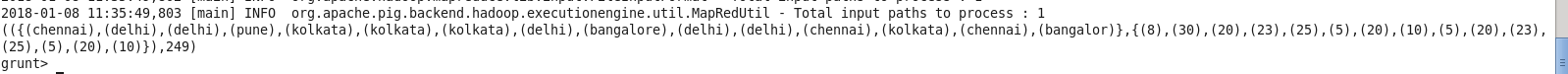
**SUM()**

**SUM()** function of Pig Latin to get the total of the numeric values of a column in a single-column bag. While computing the total, the **SUM()** function ignores the NULL values.



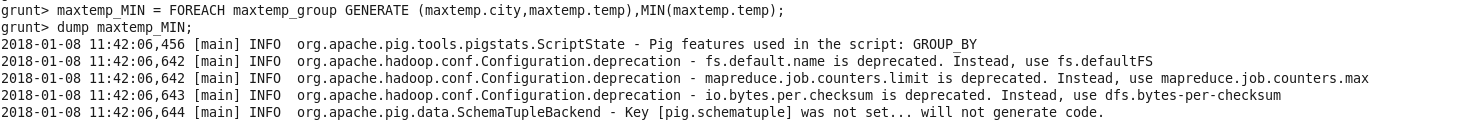






**MIN()**

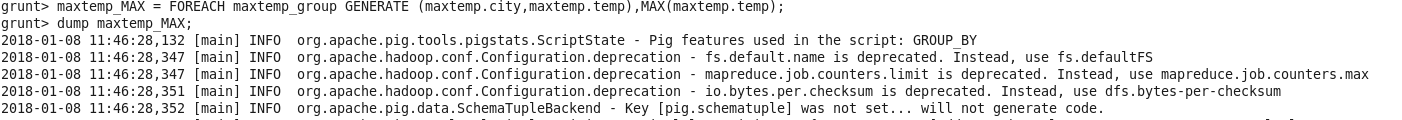
The **MIN()** function of Pig Latin is used to get the minimum (lowest) value (numeric or chararray) for a certain column in a single-column bag. While calculating the minimum value, the **MIN()** function ignores the NULL values.





**MAX()**

The Pig Latin **MAX()** function is used to calculate the highest value for a column (numeric values or chararrays) in a single-column bag. While calculating the maximum value, the **Max()** function ignores the NULL values.

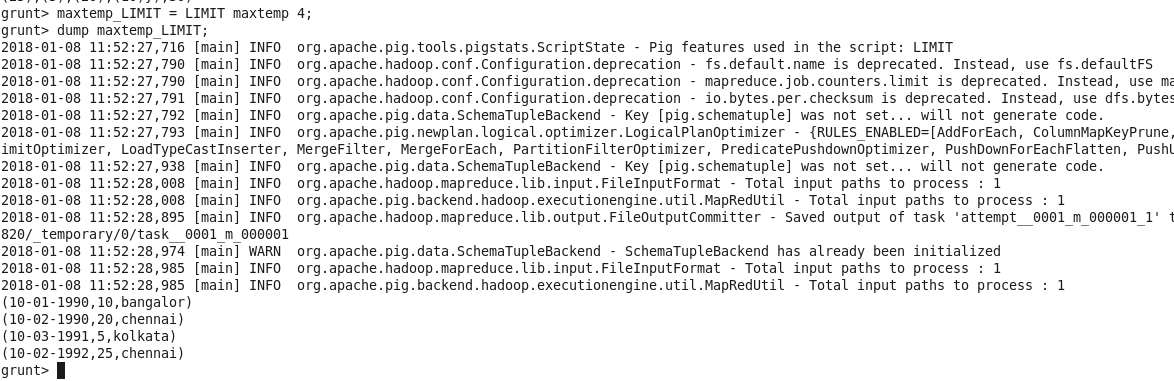




**LIMIT:**

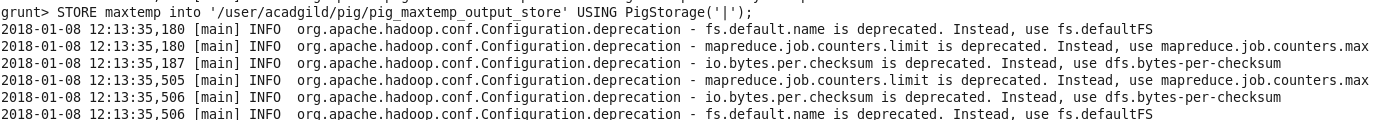
The **LIMIT** operator is used to get a limited number of tuples from a relation.

1. Get the first Four tuples in the dataset. (LIMIT maxtemp 4).



**STORE:**

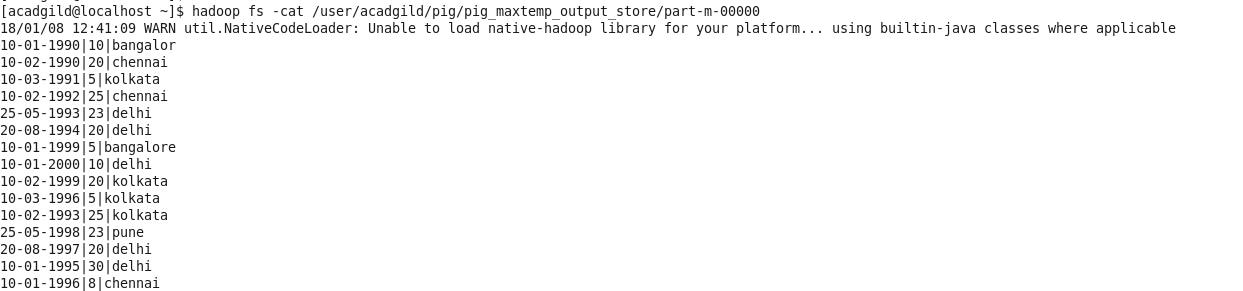
We can store the loaded data in the file system using the **store** operator.



Storage location:

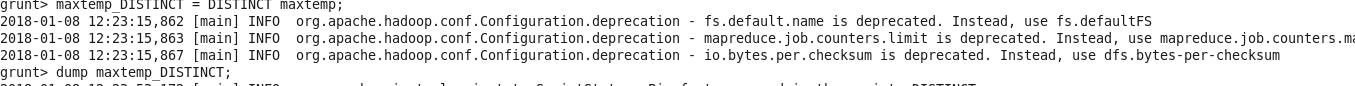


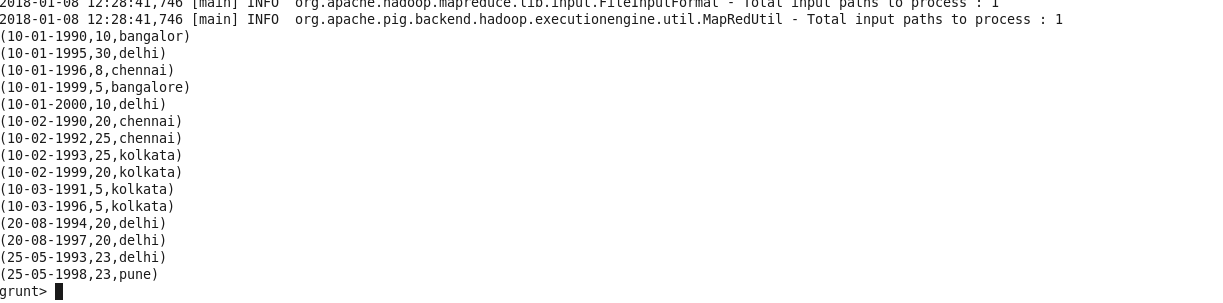
if we cat the file :



**DISTINCT:**

The **DISTINCT** operator is used to remove redundant (duplicate) tuples from a relation.

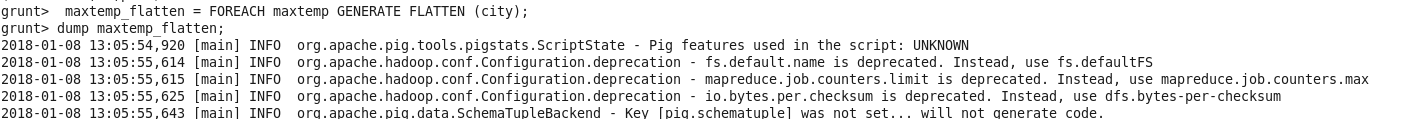


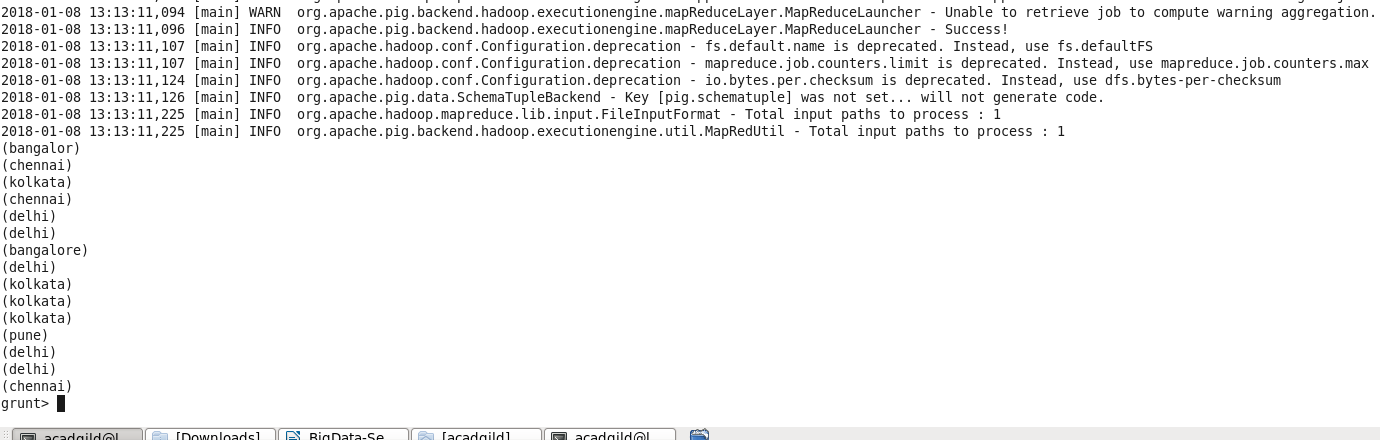


**FLATTEN:**

The **FLATTEN** operator looks like a UDF syntactically, but it is actually an operator that changes the structure of tuples and bags in a way that a UDF cannot.

Sometimes there is data in a tuple or bag and if we want to remove the level of nesting from that data then Flatten modifier in Pig can be used





**ISEMPTY:**

**IsEmpty()** function of Pig Latin is used to check if a bag or map is empty.

creating a co group and then use the empty method top of it

