

**LET US USE PIG TO DO SOME REAL LIFE
DATA MANIPULATION**

LET US USE PIG TO DO SOME REAL LIFE
DATA MANIPULATION

SUPPOSE WE ARE GIVEN LOGS FROM A SERVER

SUPPOSE WE ARE GIVEN LOGS FROM A SERVER

```
[2016-04-01 T 00:00:00.000+00:00] 2222 Error Hadoopcluster1 p121 t123 module1 error_type-1 error-message-1
[2016-04-01 T 00:00:00.000+00:00] 2224 Error Hadoopcluster3 p123 t127 module1 error_type-2 error-message-2
[2016-04-01 T 00:00:00.000+00:00] 2223 Error Hadoopcluster2 p122 t125 module1 error_type-3 error-message-3
[2016-04-01 T 00:00:00.000+00:00] message "Hi, How are you?"
[2016-04-01 T 00:01:00.000+00:00] 2222 warning Hadoopcluster1 p121 t123 module1 warning_type_1
[2016-04-01 T 00:02:00.000+00:00] 2223 warning Hadoopcluster2 p122 t125 module1 warning_type_2
[2016-04-01 T 00:02:50.000+00:00] 2224 warning Hadoopcluster3 p123 t127 module1 warning_type_3
[2016-04-01 T 00:02:00.000+00:00] 2222 Error Hadoopcluster1 p121 t123 module1 error_type-3 error-message-4
[2016-04-01 T 00:02:00.000+00:00] 2224 Error Hadoopcluster3 p123 t127 module1 error_type-2 error-message-5
[2016-04-01 T 00:02:00.000+00:00] 2223 Error Hadoopcluster2 p122 t125 module1 error_type-1 error-message-6
```

THE LOGS LOOK LIKE THIS

THE LOGS LOOK LIKE THIS

```
[2016-04-01 T 00:00:00.000+00:00] 2222 Error Hadoopcluster1 p121 t123 module1 error_type-1 error-message-1
[2016-04-01 T 00:00:00.000+00:00] 2224 Error Hadoopcluster3 p123 t127 module1 error_type-2 error-message-2
[2016-04-01 T 00:00:00.000+00:00] 2223 Error Hadoopcluster2 p122 t125 module1 error_type-3 error-message-3
[2016-04-01 T 00:00:00.000+00:00] message "Hi, How are you?"
[2016-04-01 T 00:01:00.000+00:00] 2222 warning Hadoopcluster1 p121 t123 module1 warning_type_1
[2016-04-01 T 00:02:00.000+00:00] 2223 warning Hadoopcluster2 p122 t125 module1 warning_type_2
[2016-04-01 T 00:02:50.000+00:00] 2224 warning Hadoopcluster3 p123 t127 module1 warning_type_3
[2016-04-01 T 00:02:00.000+00:00] 2222 Error Hadoopcluster1 p121 t123 module1 error_type-3 error-message-4
[2016-04-01 T 00:02:00.000+00:00] 2224 Error Hadoopcluster3 p123 t127 module1 error_type-2 error-message-5
[2016-04-01 T 00:02:00.000+00:00] 2223 Error Hadoopcluster2 p122 t125 module1 error_type-1 error-message-6
```

IT'S COMPLETELY UNSTRUCTURED... OR IS IT?

IF YOU LOOK CLOSELY, YOU CAN SEE SOME PATTERNS

[2016-04-01	T	00:00:00.000+00:00]	2222	Error	Hadoopcluster1	p121	t123	module1	error_type-1	error-message-1
[2016-04-01	T	00:00:00.000+00:00]	2224	Error	Hadoopcluster3	p123	t127	module1	error_type-2	error-message-2
[2016-04-01	T	00:00:00.000+00:00]	2223	Error	Hadoopcluster2	p122	t125	module1	error_type-3	error-message-3
[2016-04-01	T	00:00:00.000+00:00]		message "Hi, How are you?"						
[2016-04-01	T	00:01:00.000+00:00]	2222	warning	Hadoopcluster1	p121	t123	module1	warning_type_1	
[2016-04-01	T	00:02:00.000+00:00]	2223	warning	Hadoopcluster2	p122	t125	module1	warning_type_2	
[2016-04-01	T	00:02:50.000+00:00]	2224	warning	Hadoopcluster3	p123	t127	module1	warning_type_3	
[2016-04-01	T	00:02:00.000+00:00]	2222	Error	Hadoopcluster1	p121	t123	module1	error_type-3	error-message-4
[2016-04-01	T	00:02:00.000+00:00]	2224	Error	Hadoopcluster3	p123	t127	module1	error_type-2	error-message-5
[2016-04-01	T	00:02:00.000+00:00]	2223	Error	Hadoopcluster2	p122	t125	module1	error_type-1	error-message-6

IF YOU LOOK CLOSELY, YOU CAN SEE SOME PATTERNS

[2016-04-01	T	00:00:00.000+00:00]	2222	Error	Hadoopcluster1	p121	t123	module1	error_type-1	error-message-1
[2016-04-01	T	00:00:00.000+00:00]	2224	Error	Hadoopcluster3	p123	t127	module1	error_type-2	error-message-2
[2016-04-01	T	00:00:00.000+00:00]	2223	Error	Hadoopcluster2	p122	t125	module1	error_type-3	error-message-3
[2016-04-01	T	00:00:00.000+00:00]	message "Hi, How are you?"							
[2016-04-01	T	00:01:00.000+00:00]	2222	warning	Hadoopcluster1	p121	t123	module1	warning_type_1	
[2016-04-01	T	00:02:00.000+00:00]	2223	warning	Hadoopcluster2	p122	t125	module1	warning_type_2	
[2016-04-01	T	00:02:50.000+00:00]	2224	warning	Hadoopcluster3	p123	t127	module1	warning_type_3	
[2016-04-01	T	00:02:00.000+00:00]	2222	Error	Hadoopcluster1	p121	t123	module1	error_type-3	error-message-4
[2016-04-01	T	00:02:00.000+00:00]	2224	Error	Hadoopcluster3	p123	t127	module1	error_type-2	error-message-5
[2016-04-01	T	00:02:00.000+00:00]	2223	Error	Hadoopcluster2	p122	t125	module1	error_type-1	error-message-6

TimeStamp(\$0 , \$2)

Server_Number
(\$3)

TYPE_OF_MESSAGE
(\$4)

HADOOP_CLUSTER
(\$5)

PROCESS_ID
(\$6)

THREAD_ID
(\$7)

MODULE
(\$8)

ERROR_MESSAGES
(\$10)

TYPE_OF_ERROR
(\$9)

```
[2016-04-01 T 00:00:00.000+00:00] 2222 Error Hadoopcluster1 p121 t123 module1 error_type-1 error-message-1
[2016-04-01 T 00:00:00.000+00:00] 2224 Error Hadoopcluster3 p123 t127 module1 error_type-2 error-message-2
[2016-04-01 T 00:00:00.000+00:00] 2223 Error Hadoopcluster2 p122 t125 module1 error_type-3 error-message-3
[2016-04-01 T 00:00:00.000+00:00] message "Hi, How are you?"
[2016-04-01 T 00:01:00.000+00:00] 2222 warning Hadoopcluster1 p121 t123 module1 warning_type_1
[2016-04-01 T 00:02:00.000+00:00] 2223 warning Hadoopcluster2 p122 t125 module1 warning_type_2
[2016-04-01 T 00:02:50.000+00:00] 2224 warning Hadoopcluster3 p123 t127 module1 warning_type_3
[2016-04-01 T 00:02:00.000+00:00] 2222 Error Hadoopcluster1 p121 t123 module1 error_type-3 error-message-4
[2016-04-01 T 00:02:00.000+00:00] 2224 Error Hadoopcluster3 p123 t127 module1 error_type-2 error-message-5
[2016-04-01 T 00:02:00.000+00:00] 2223 Error Hadoopcluster2 p122 t125 module1 error_type-1 error-message-6
```

TYPE_OF_MESSAGE

(\$4)

ERROR

MESSAGE

WARNING

LOGS CONTAIN THREE TYPES OF MESSAGES

ERROR

```
[2016-04-01 T 00:00:00.000+00:00] 2222 Error Hadoopcluster1 p121 t123 module1 error_type-1 error-message-1
[2016-04-01 T 00:00:00.000+00:00] 2224 Error Hadoopcluster3 p123 t127 module1 error_type-2 error-message-2
[2016-04-01 T 00:00:00.000+00:00] 2223 Error Hadoopcluster2 p122 t125 module1 error_type-3 error-message-3
```

MESSAGE

```
[2016-04-01 T 00:00:00.000+00:00] message "Hi, How are you?"
```

WARNING

```
[2016-04-01 T 00:01:00.000+00:00] 2222 warning Hadoopcluster1 p121 t123 module1 warning_type_1
[2016-04-01 T 00:02:00.000+00:00] 2223 warning Hadoopcluster2 p122 t125 module1 warning_type_2
[2016-04-01 T 00:02:50.000+00:00] 2224 warning Hadoopcluster3 p123 t127 module1 warning_type_3
```

WHAT IF WE WANT SOME USEFUL SUMMARY INFORMATION?

ERROR

```
[2016-04-01 T 00:00:00.000+00:00] 2222 Error Hadoopcluster1 p121 t123 module1 error_type-1 error-message-1  
[2016-04-01 T 00:00:00.000+00:00] 2224 Error Hadoopcluster3 p123 t127 module1 error_type-2 error-message-2  
[2016-04-01 T 00:00:00.000+00:00] 2223 Error Hadoopcluster2 p122 t125 module1 error_type-3 error-message-3
```

MESSAGE

```
[2016-04-01 T 00:00:00.000+00:00] message "Hi, How are you?"
```

WARNING

```
[2016-04-01 T 00:01:00.000+00:00] 2222 warning Hadoopcluster1 p121 t123 module1 warning_type_1  
[2016-04-01 T 00:02:00.000+00:00] 2223 warning Hadoopcluster2 p122 t125 module1 warning_type_2  
[2016-04-01 T 00:02:50.000+00:00] 2224 warning Hadoopcluster3 p123 t127 module1 warning_type_3
```

DUMP OF ERROR DATA

SUMMARY OF ERROR
DATA

DUMP OF WARNING
DATA

WE WILL GET RID OF MESSAGES

WHAT IF WE WANT SOME USEFUL SUMMARY INFORMATION?

ERROR

```
[2016-04-01T00:00:00.000+00:00] 2222 Error Hadoopcluster1 p121 t123 module1 error_type-1 error-message-1  
[2016-04-01T00:00:00.000+00:00] 2224 Error Hadoopcluster3 p123 t127 module1 error_type-2 error-message-2  
[2016-04-01T00:00:00.000+00:00] 2223 Error Hadoopcluster2 p122 t125 module1 error_type-3 error-message-3
```

DUMP OF ERROR DATA

SIMPLY DUMPS ALL THE ERROR DATA

Time Stamp	Month	Day	Server_Number	Error_Type	Hadoop_Cluster	Process ID	Thread ID	Module	Type of Error	Error_Messages
2016-01-01T00:00:00.000+05:30	1	1	2222	Error	Hadoopcluster1	p121	t123	module1	error_type-1	error-message-1
2016-01-01T00:00:00.000+05:30	1	1	2224	Error	Hadoopcluster3	p123	t127	module1	error_type-2	error-message-2
2016-01-01T00:00:00.000+05:30	1	1	2223	Error	Hadoopcluster2	p122	t125	module1	error_type-3	error-message-3
2016-01-01T00:02:00.000+05:30	1	1	2222	Error	Hadoopcluster1	p121	t123	module1	error_type-3	error-message-4
2016-01-01T00:02:00.000+05:30	1	1	2224	Error	Hadoopcluster3	p123	t127	module1	error_type-2	error-message-5
2016-01-01T00:02:00.000+05:30	1	1	2223	Error	Hadoopcluster2	p122	t125	module1	error_type-1	error-message-6

WHAT IF WE WANT SOME USEFUL SUMMARY INFORMATION?

ERROR

```
[2016-04-01 T 00:00:00.000+00:00] 2222 Error Hadoopcluster1 p121 t123 module1 error_type-1 error-message-1  
[2016-04-01 T 00:00:00.000+00:00] 2224 Error Hadoopcluster3 p123 t127 module1 error_type-2 error-message-2  
[2016-04-01 T 00:00:00.000+00:00] 2223 Error Hadoopcluster2 p122 t125 module1 error_type-3 error-message-3
```

MESSAGES

```
[2016-04-01 T 00:00:00.000+00:00] message "Hi, How are you?"
```

WARNING

```
[2016-04-01 T 00:01:00.000+00:00] 2222 warning Hadoopcluster1 p121 t123 module1 warning_type_1  
[2016-04-01 T 00:02:00.000+00:00] 2223 warning Hadoopcluster2 p122 t125 module1 warning_type_2  
[2016-04-01 T 00:02:50.000+00:00] 2224 warning Hadoopcluster3 p123 t127 module1 warning_type_3
```

DUMP OF ERROR DATA

SUMMARY OF ERROR
DATA

DUMP OF WARNING
DATA

WE WILL GET RID OF MESSAGES

WHAT IF WE WANT SOME USEFUL SUMMARY INFORMATION?

ERROR								
[2016-04-01T00:00:00.000+00:00]	2222	Error	Hadoopcluster1	p121	t123	module1	error_type-1	error-message-1
[2016-04-01T00:00:00.000+00:00]	2224	Error	Hadoopcluster3	p123	t127	module1	error_type-2	error-message-2
[2016-04-01T00:00:00.000+00:00]	2223	Error	Hadoopcluster2	p122	t125	module1	error_type-3	error-message-3

SUMMARY OF ERROR DATA

IS A SUMMARY OF ERRORS ON SERVERS

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module1	1	error-message-4
2223	error_type-1	2016-01-01T00:02:00.000+05:30	p122	t125	module1	1	error-message-6
2223	error_type-3	2016-01-01T00:00:00.000+05:30	p122	t125	module1	1	error-message-3
2224	error_type-2	2016-01-01T00:02:00.000+05:30	p123	t127	module1	2	error-message-5,error-message-2

SUMMARY OF ERROR DATA

IS A SUMMARY OF SERVER_NUMBER AND ERRORS ON THEM

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module 1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module 1	1	error-message-4
2223	error_type-1	2016-01-01T00:02:00.000+05:30	p122	t125	module 1	1	error-message-6
2223	error_type-3	2016-01-01T00:00:00.000+05:30	p122	t125	module 1	1	error-message-3
2224	error_type-2	2016-01-01T00:02:00.000+05:30	p123	t127	module 1	2	error-message-5,error-message-2

WE USE THIS TABLE TO GENERATE VARIOUS ERROR STATISTICS FOR SERVERS

USING GROUP BY

SUMMARY OF ERROR DATA

IS A SUMMARY OF SERVER_NUMBER AND ERRORS ON THEM

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module 1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module 1	1	error-message-4
2223	error_type-1	2016-01-01T00:02:00.000+05:30	p122	t125	module 1	1	error-message-6
2223	error_type-3	2016-01-01T00:00:00.000+05:30	p122	t125	module 1	1	error-message-3
2224	error_type-2	2016-01-01T00:02:00.000+05:30	p123	t127	module 1	2	error-message-5,error-message-2

THESE THREE COLUMNS ARE SIMPLY TAKEN FROM THE
ERROR_DATA

SUMMARY OF ERROR DATA

IS A SUMMARY OF SERVER NUMBER AND ERRORS ON THEM

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module 1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module 1	1	error-message-4
2223	error_type-1	2016-01-01T00:02:00.000+05:30	p122	t125	module 1	1	error-message-6
2223	error_type-3	2016-01-01T00:00:00.000+05:30	p122	t125	module 1	1	error-message-3
2224	error_type-2	2016-01-01T00:02:00.000+05:30	p123	t127	module 1	2	error-message-5,error-message-2

THIS IS THE COLUMN THAT CAPTURES THE LAST TIME STAMP OF AN ERROR ON A SERVER

SUMMARY OF ERROR DATA

IS A SUMMARY OF SERVER_NUMBER AND ERRORS ON THEM

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module 1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module 1	1	error-message-4
2223	error_type-1	2016-01-01T00:02:00.000+05:30	p122	t125	module 1	1	error-message-6
2223	error_type-3	2016-01-01T00:00:00.000+05:30	p122	t125	module 1	1	error-message-3
2224	error_type-2	2016-01-01T00:02:00.000+05:30	p123	t127	module 1	2	error-message-5,error-message-2

THESE TWO COLUMNS ARE GENERATED USING AGGREGATE FUNCTIONS

SUMMARY OF ERROR DATA

IS A SUMMARY OF SERVER_NUMBER AND ERRORS ON THEM

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module 1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module 1	1	error-message-4
2223	error_type-1	2016-01-01T00:02:00.000+05:30	p122	t125	module 1	1	error-message-6
2223	error_type-3	2016-01-01T00:00:00.000+05:30	p122	t125	module 1	1	error-message-3
2224	error_type-2	2016-01-01T00:02:00.000+05:30	p123	t127	module 1	2	error-message-5,error-message-2

THIS COLUMN TELLS US THE NUMBER OF TIMES A PARTICULAR ERROR OCCURRED ON EACH SERVERS

SUMMARY OF ERROR DATA

IS A SUMMARY OF SERVER_NUMBER AND ERRORS ON THEM

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module 1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module 1	1	error-message-4
2223	error_type-1	2016-01-01T00:02:00.000+05:30	p122	t125	module 1	1	error-message-6
2223	error_type-3	2016-01-01T00:00:00.000+05:30	p122	t125	module 1	1	error-message-3
2224	error_type-2	2016-01-01T00:02:00.000+05:30	p123	t127	module 1	2	error-message-5,error-message-2

THIS COLUMN AGGREGATES ALL THE ERROR MESSAGES IN A SINGLE FIELD

WHAT IF WE WANT SOME USEFUL SUMMARY INFORMATION?

ERROR

```
[2016-04-01 T 00:00:00.000+00:00] 2222 Error Hadoopcluster1 p121 t123 module1 error_type-1 error-message-1  
[2016-04-01 T 00:00:00.000+00:00] 2224 Error Hadoopcluster3 p123 t127 module1 error_type-2 error-message-2  
[2016-04-01 T 00:00:00.000+00:00] 2223 Error Hadoopcluster2 p122 t125 module1 error_type-3 error-message-3
```

MESSAGES

```
[2016-04-01 T 00:00:00.000+00:00] message "Hi, How are you?"
```

WARNING

```
[2016-04-01 T 00:01:00.000+00:00] 2222 warning Hadoopcluster1 p121 t123 module1 warning_type_1  
[2016-04-01 T 00:02:00.000+00:00] 2223 warning Hadoopcluster2 p122 t125 module1 warning_type_2  
[2016-04-01 T 00:02:50.000+00:00] 2224 warning Hadoopcluster3 p123 t127 module1 warning_type_3
```

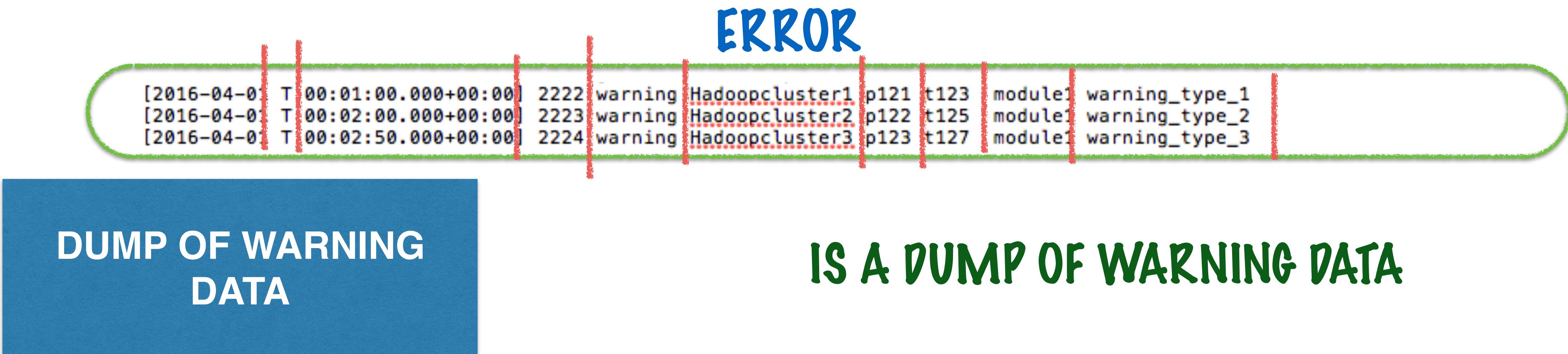
DUMP OF ERROR DATA

SUMMARY OF ERROR
DATA

DUMP OF WARNING
DATA

WE WILL GET RID OF MESSAGES

WHAT IF WE WANT SOME USEFUL SUMMARY INFORMATION?



Time Stamp	Month	Day	Server_Number	Error_Type	Hadoop_Cluster	Process ID	Thread ID	Module	Type of Error	Error_Messages
2016-01-01T00:01:00.000+05:30	1	1		2222 warning	Hadoopcluster1	p121	t123	module1	warning_type_1	
2016-01-01T00:02:00.000+05:30	1	1		2223 warning	Hadoopcluster2	p122	t125	module1	warning_type_2	
2016-01-01T00:02:50.000+05:30	1	1		2224 warning	Hadoopcluster3	p123	t127	module1	warning_type_3	

THIS COLUMN WILL BE BLANK, THERE ARE NO WARNING MESSAGES

WE NEED A SCRIPT THAT WILL GENERATE THESE THREE RELATIONS

DUMP OF ERROR DATA

SUMMARY OF ERROR
DATA

DUMP OF WARNING
DATA

THE SCRIPT THAT WILL GENERATE THESE THREE RELATIONS

```
grunt> log_data = load 'log_data.txt' using PigStorage(' ');
grunt> log_data_1 = filter log_data by not $3 == 'message';
grunt> log_data_2= foreach log_data_1 generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data_3 = foreach log_data_2 generate CONCAT($0,' '),$2,$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data_4 = foreach log_data_3 generate CONCAT($0,$1),$2,$3,$4,$5,$6,$7,$8,$9;
grunt> log_data_5 = foreach log_data_4 generateToDate($0,'yyyy-mm-dd HH:mm:ss'),$1,$2,$3,$4,$5,$6,$7,$8;
grunt> log_data_6 = foreach log_data_5 generate $0 as Time_Stamp,GetMonth($0) as Month,GetDay($0) as Day,$1 as
server_number,$2 as type_of_message,$3 as cluster_name,$4 as process_ID,$5 as thread_ID ,$6 as module,$7 as error_type
,$8 as error_message;

grunt> split log_data_6 into warnings_data if type_of_message == 'warning', error_data if type_of_message == 'Error';
grunt> groupd = group error_data by (server_number,error_type);

grunt> error_summary = foreach groupd{
grunt> number_of_occurrences = COUNT(error_data);
grunt> error_messages = BagToString(error_data.error_message,',');
grunt> error_data = foreach error_data generate server_number, error_type, Time_Stamp, process_ID, thread_ID, module;
grunt> sorted_error_messages = order error_data by Time_Stamp desc;
grunt> top_item = limit sorted_error_messages 1;
grunt> generate flatten(top_item),number_of_occurrences,error_messages;
grunt> };

grunt> store error_summary into 'error_summary';
grunt> store error_data into 'error_data';
grunt> store warnings_data into 'warnings_data';
```

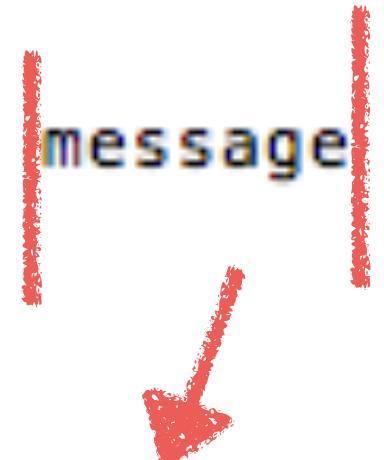
```
grunt> log_data = load 'log_data.txt' using PigStorage(' ');
grunt> log_data_1 = filter log_data by not $3 == 'message';
grunt> log_data = foreach log_data generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,' '),$2,$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,$1),$2,$3,$4,$5,$6,$7,$8,$9;
grunt> log_data = foreach log_data generateToDate($0,'yyyy-mm-dd HH:mm:ss'),$1,$2,$3,$4,$5,$6,$7,$8;
grunt> log_data = foreach log_data generate $0 as Time_Stamp,GetMonth($0) as Month,GetDay($0) as Day,$1 as
server_number,$2 as type_of_message,$3 as cluster_name,$4 as process_ID,$5 as thread_ID,$6 as module,$7 as error_type
,$8 as error_message;
```

**THIS WILL HELP US IN FILTERING OUT
MESSAGE LOGS**

```
grunt> error_summary = foreach groupd{
grunt>   number_of_occurrences = COUNT(error_data);
grunt>   error_messages = BagToString(error_data.error_message,',');
grunt>   error_data = foreach error_data [2016-04-01 T 00:00:00.000+00:00] message "Hi, How are you?" | dule;
grunt>   sorted_error_messages = order error_data by Time_Stamp desc;
grunt>   top_item = limit sorted_error_messages 1;
grunt>   generate flatten(top_item),number_of_occurrences,error_messages;
grunt> };
```

```
grunt> store error_summary into 'error_summary';
grunt> store error_data into 'error_data';
grunt> store warnings_data into 'warnings_data';
```

\$3 == 'message'



```
grunt> log_data = load 'log_data.txt' using PigStorage(' ');
grunt> log_data = filter log_data by not $3 == 'message';
grunt> log_data_2 = foreach log_data_1 generate SUBSTRING($0,1,11),
$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0, ' '),
$2,$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,$1),
$2,$3,$4,$5,$6,$7,$8,$9;
grunt> log_data = foreach log_data generateToDate($0,'yyyy-mm-dd HH:mm:ss'),
$1,$2,$3,$4,$5,$6,$7,$8;
grunt> log_data = foreach log_data generate $0 as Time_Stamp,GetMonth($0) as Month,GetDay($0) as Day,$1 as
server_number,$2 as type_of_message,$3 as cluster,$4 as process_ID,$5 as thread_ID,$6 as module,$7 as error_type
,$8 as error_message;
```

**WE GET RID OF OPENING BRACKET AND CLOSING
BRACKET FROM FIELD 1 AND FIELD 3**

```
grunt> split log_data into warnings_data if type_of_message == 'warning', error_data if type_of_message == 'Error';
grunt> copyToLocal(error_data, 'error_data');
[2016-04-01 |T| 00:00:00.000+00:00] 2222 Error Hadoopcluster1 p121 t123 module1 error_type-
[2016-04-01 |T| 00:00:00.000+00:00] 2224 Error Hadoopcluster3 p123 t127 module1 error_type-
[2016-04-01 |T| 00:00:00.000+00:00] 2223 Error Hadoopcluster2 p122 t125 module1 error_type-
grunt> sorted_error_messages = order error_data by Time_Stamp desc;
grunt> top_item = limit sorted_error_messages 1;
grunt> generate flatten(top_item),number_of_occurrences,error_messages;
grunt> };

grunt> store error_summary into 'error_summary';
grunt> store error_data into 'error_data';
grunt> store warnings data into 'warnings data';
```

```
grunt> log_data = load 'log_data.txt' using PigStorage(' ');
grunt> log_data = filter log_data by not $3 == 'message';
grunt> log_data_2 = foreach log_data_1 generate SUBSTRING($0,1,11),
$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0, ' '),
$2,$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,$1),
$2,$3,$4,$5,$6,$7,$8,$9;
grunt> log_data = foreach log_data generateToDate($0,'yyyy-mm-dd HH:mm:ss'),
$1,$2,$3,$4,$5,$6,$7,$8;
grunt> log_data = foreach log_data generate $0 as Time_Stamp,GetMonth($0) as Month,GetDay($0) as Day,$1 as
server_number,$2 as type_of_message,$3 as cluster_name,$4 as process_ID,$5 as thread_ID ,$6 as module,$7 as error_type
,$8 as error_message;
```

THE DATA LOOKS LIKE THIS AFTER THIS COMMAND

```
grunt> split log_data into warnings_data if type_of_message == 'warning', error_data if type_of_message == 'Error';
grunt> group = group error_data by (server_number,error_time);
(2016-04-01,T,00:00:00,2222>Error,Hadoopcluster1,p121,t123,module1,error_type-1,error-message-1)
(2016-04-01,T,00:00:00,2224>Error,Hadoopcluster3,p123,t127,module1,error_type-2,error-message-2)
(2016-04-01,T,00:00:00,2223>Error,Hadoopcluster2,p122,t125,module1,error_type-3,error-message-3)
(2016-04-01,T,00:01:00,2222,warning,Hadoopcluster1,p121,t123,module1,warning_type_1, )
(2016-04-01,T,00:02:00,2223,warning,Hadoopcluster2,p122,t125,module1,warning_type_2, )
grunt> top_item = limit sorted_error_messages 1;
grunt> generate flatten(top_item),number_of_occurrences,error_messages;
grunt> };

grunt> store error_summary into 'error_summary';
grunt> store error_data into 'error_data';
grunt> store warnings_data into 'warnings_data';
```

```
grunt> log_data = load 'log_data.txt' using PigStorage(' ');
grunt> log_data = filter log_data by not $3 == 'message';
grunt> log_data_3 = foreach log_data generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),
$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data_3 = foreach log_data_2 generate CONCAT($0, ' '),
$2,$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data_4 = foreach log_data_3 generate CONCAT($0,$1),
$2,$3,$4,$5,$6,$7,$8,$9;
grunt> log_data = foreach log_data generate ToDate($0,'yyyy-mm-dd HH:mm:ss'),
$1,$2,$3,$4,$5,$6,$7,$8;
grunt> log_data = foreach log_data generate $0 as Time_Stamp,GetMonth($0) as Month,GetDay($0) as Day,$1 as
server_number,$2 as type_of_message,$3 as
,$8 as error_message;
```

THESE TWO STATEMENTS ARE USED TO CONCATENATE FIELD 0 AND FIELD 2 WITH '' IN BETWEEN

```
grunt> split log_data into warnings_data if type_of_message == 'warning', error_data if type_of_message == 'Error';
grunt> groupd = group error_data by (server_number,error_type);
```

```
grunt> error_summary = foreach groupd{
grn (2016-04-01,T,00:00:00,2222>Error,Hadoopcluster1,p121,t123,module1,error_type-1,error-mess
grn (2016-04-01,T,00:00:00,2224>Error,Hadoopcluster3,p123,t127,module1,error_type-2,error-mess
grn (2016-04-01,T,00:00:00,2223>Error,Hadoopcluster2,p122,t125,module1,error_type-3,error-mess
grn (2016-04-01,T,00:01:00,2222>warning,Hadoopcluster1,p121,t123,module1,warning_type_1,
grn (2016-04-01,T,00:02:00,2223>warning,Hadoopcluster2,p122,t125,module1,warning_type_2,
```

```
grunt> store error_summary into 'error_summary';
grunt> store error_data into 'error_data';
grunt> store warnings_data into 'warnings_data';
```

THE SCRIPT THAT WILL GENERATE THESE THREE TABLES IS

```
grunt> log_data = load 'log_data.txt' using PigStorage(' ');
grunt> log_data = filter log_data by not $3 == 'message';
grunt> log_data = foreach log_data generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data_3 = foreach log_data_2 generate CONCAT($0,' '),
$2,$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data_4 = foreach log_data_3 generate CONCAT($0,$1),
$2,$3,$4,$5,$6,$7,$8,$9;
grunt> log_data = foreach log_data generate ToDate($0,'yyyy-mm-dd HH:mm:ss'),
$1,$2,$3,$4,$5,$6,$7,$8;
grunt> log_data = foreach log_data generate $0 as Time_Stamp,GetMonth($0) as Month,GetDay($0) as Day,$1 as
server_number,$2 as type_of_message,$3 as cluster_name,$4 as process_ID,$5 as thread_ID,$6 as module,$7 as error_type
,$8 as error_message;
```

log_data LOOKS CLEANER NOW

```
grunt> split log_data into warnings_data if type_of_message == 'warning', error_data if type_of_message == 'Error';
grunt> groupd = group error_data by (server_number,error_type);
```

```
grunt> (2016-04-01 00:00:00,2222>Error,Hadoopcluster1,p121,t123,module1,error_type-1,error-message-1)
grunt> (2016-04-01 00:00:00,2224>Error,Hadoopcluster3,p123,t127,module1,error_type-2,error-message-2)
grunt> (2016-04-01 00:00:00,2223>Error,Hadoopcluster2,p122,t125,module1,error_type-3,error-message-3)
grunt> (2016-04-01 00:01:00,2222,warning,Hadoopcluster1,p121,t123,module1,warning_type_1, )
grunt> (2016-04-01 00:02:00,2223,warning,Hadoopcluster2,p122,t125,module1,warning_type_2, )
grunt> generate flatten(top_item),number_of_occurrences,error_messages;
grunt> 
```

```
grunt> store error_summary into 'error_summary';
grunt> store error_data into 'error_data';
grunt> store warnings_data into 'warnings_data';
```

```
grunt> log_data = load 'log_data.txt' using PigStorage(' ');
grunt> log_data = filter log_data by not $3 == 'message';
grunt> log_data = foreach log_data generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),
$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,' '),$2,$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,$1),$2,$3,$4,$5,$6,$7,$8,$9;
grunt> log_data_5 = foreach log_data_4 generate ToDate($0, 'yyyy-mm-dd HH:mm:ss'),$1,$2,$3,$4,$5,$6,$7,$8;
```

WE CAST THE FIRST FIELD INTO DATETIME DATA TYPE

```
(2016-04-01 00:00:00,2222>Error,Hadoopcluster1,p121,t123,module1,error_type-1,error-message-1)
(2016-04-01 00:00:00,2224>Error,Hadoopcluster3,p123,t127,module1,error_type-2,error-message-2)
(2016-04-01 00:00:00,2223>Error,Hadoopcluster2,p122,t125,module1,error_type-3,error-message-3)
(2016-04-01 00:01:00,2222,warning,Hadoopcluster1,p121,t123,module1,warning_type_1, )
(2016-04-01 00:02:00,2223,warning,Hadoopcluster2,p122,t125,module1,warning_type_2, )
```

Success.

```
grunt> describe error_summary
error_summary: {id: string, type: string, cluster: string, host: string, timestamp: string, module: string, error_type: string, warning_type: string}
```

```
grunt> store error_summary into error_data;
```

```
grunt> store warning_summary into warning_data;
```

```
grunt> log_data = load 'log_data.txt' using PigStorage(' ');
grunt> log_data = filter log_data by not $3 == 'message';
grunt> log_data = foreach log_data generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,' '),$2,$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,$1),$2,$3,$4,$5,$6,$7,$8,$9;
grunt> log_data = foreach log_data generate ToDate($0,'yyyy-mm-dd HH:mm:ss'),$1,$2,$3,$4,$5,$6,$7,$8;
grunt> log_data_6 = foreach log_data_5 generate $0 as Time_Stamp,GetMonth($0)
as Month,GetDay($0) as Day,$1 as server_number,$2 as type_of_message,$3 as
cluster_name,$4 as process_ID,$5 as thread_ID ,$6 as module,$7 as error_type ,
$8 as error_message;
```

```
grunt> split log_data into warnings_data if type_of_message == 'warning', error_data if type_of_message == 'Error';
grunt> groupd = group error_data by (server_number,error_type);
```

```
grunt> error_summary = foreach groupd{
grunt> number_of_occurrences = COUNT(error_data);
grunt> error_messages = BagToString(error_data.error_message);
grunt> error_data = foreach error_data generate server_number,error_type,$1,$2,$3,$4,$5,$6,$7,$8,$9,$10;
grunt> sorted_error_messages = order error_data by Time_Stamp DESC;
grunt> top_item = limit sorted_error_messages 1;
grunt> generate flatten(top_item),number_of_occurrences,error_messages;
grunt> };
```

```
grunt> store error_summary into 'error_summary';
grunt> store error_data into 'error_data';
grunt> store warnings_data into 'warnings_data';
```

**AT THIS STEP WE NAME VARIOUS FIELDS
AND SPECIFY THE SCHEMA**

[2016-04-01	T	00:00:00.000+00:00]	2222	Error	Hadoopcluster1	p121	t123	module1	error_type-1	error-message-1
[2016-04-01	T	00:00:00.000+00:00]	2224	Error	Hadoopcluster3	p123	t127	module1	error_type-2	error-message-2
[2016-04-01	T	00:00:00.000+00:00]	2223	Error	Hadoopcluster2	p122	t125	module1	error_type-3	error-message-3
[2016-04-01	T	00:00:00.000+00:00]	message "Hi, How are you?"							
[2016-04-01	T	00:01:00.000+00:00]	2222	warning	Hadoopcluster1	p121	t123	module1	warning_type_1	
[2016-04-01	T	00:02:00.000+00:00]	2223	warning	Hadoopcluster2	p122	t125	module1	warning_type_2	
[2016-04-01	T	00:02:50.000+00:00]	2224	warning	Hadoopcluster3	p123	t127	module1	warning_type_3	
[2016-04-01	T	00:02:00.000+00:00]	2222	Error	Hadoopcluster1	p121	t123	module1	error_type-3	error-message-4
[2016-04-01	T	00:02:00.000+00:00]	2224	Error	Hadoopcluster3	p123	t127	module1	error_type-2	error-message-5
[2016-04-01	T	00:02:00.000+00:00]	2223	Error	Hadoopcluster2	p122	t125	module1	error_type-1	error-message-6

TimeStamp(\$0 , \$2)

Server_Number
(\$3)

TYPE_OF_MESSAGE
(\$4)

HADOOP_CLUSTER
(\$5)

PROCESS_ID
(\$6)

THREAD_ID
(\$7)

MODULE
(\$8)

TYPE_OF_ERROR
(\$9)

ERROR_MESSAGES
(\$10)

```
grunt> log_data = load 'log_data.txt' using PigStorage(' ');
grunt> log_data = filter log_data by not $3 == 'message';
grunt> log_data = foreach log_data generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,' '),$2,$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,$1),$2,$3,$4,$5,$6,$7,$8,$9;
grunt> log_data = foreach log_data generateToDate($0,'yyyy-mm-dd HH:mm:ss'),$1,$2,$3,$4,$5,$6,$7,$8;
grunt> log_data = foreach log_data generate $0 as Time_Stamp,GetMonth($0) as Month,GetDay($0) as Day,$1 as
server_number,$2 as type_of_message,$3 as cluster_name,$4 as process_ID,$5 as thread_ID ,$6 as module,$7 as
error_type ,$8 as error_message;
```

```
grunt> split log_data_6 into warnings_data if type_of_message ==
'warning', error_data if type_of_message == 'Error';
grunt> groupd = group error_data by (server_number,error_type);
```

```
grunt> error_summary = foreach groupd generate count(error_data);
grunt> number_of_occurrences = COUNT(error_summary);
grunt> error_messages = BagToString(error_data.error_message,',');
grunt> error_data = foreach error_data generate server_number,error_type,error_message,Time_Stamp,thread_ID, module;
grunt> sorted_error_messages = order error_data by Time_Stamp asc;
grunt> top_item = limit sorted_error_messages 1;
grunt> generate flatten(top_item),number_of_occurrences,error_messages;
grunt> };
```

```
grunt> store error_summary into 'error_summary';
grunt> store error_data into 'error_data';
grunt> store warnings_data into 'warnings_data';
```

**NOW WE SPLIT DATA INTO TWO RELATIONS
USING type_of_message**

```
grunt> split log_data_6 into warnings_data if type_of_message ==  
'warning', error_data if type_of_message == 'Error';
```

warnings_data WILL HAVE

DUMP OF WARNING DATA

Time Stamp	Month	Day	Server Number	Error Type	Hadoop Cluster	Process ID	Thread ID	Module	Type of Error	Error Messages
2016-01-01T00:01:00.000+05:30	1	1	2222	warning	Hadoopcluster1	p121	t123	module1	warning_type_1	
2016-01-01T00:02:00.000+05:30	1	1	2223	warning	Hadoopcluster2	p122	t125	module1	warning_type_2	
2016-01-01T00:02:50.000+05:30	1	1	2224	warning	Hadoopcluster3	p123	t127	module1	warning_type_3	

```

grunt> log_data = load 'log_data.txt' using PigStorage(' ');
grunt> log_data = filter log_data by not $3 == 'message';
grunt> log_data = foreach log_data generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,' '),$2,$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,$1),$2,$3,$4,$5,$6,$7,$8,$9;
grunt> log_data = foreach log_data generate ToDate($0,'yyyy-mm-dd HH:mm:ss'),$1,$2,$3,$4,$5,$6,$7,$8;
grunt> log_data = foreach log_data generate $0 as Time_Stamp,GetMonth($0) as Month,GetDay($0) as Day,$1 as
server_number,$2 as type_of_message,$3 as cluster_name,$4 as process_ID,$5 as thread_ID ,$6 as module,$7 as
error_type ,$8 as error_message;

```

```

grunt> split log_data into warnings_data if type_of_message == 'warning',
error_data if type_of_message == 'Error';

```

error_data WILL HAVE

DUMP OF ERROR DATA

Time Stamp	Month	Day	Server_Number	Error_Type	Hadoop_Cluster	Process ID	Thread ID	Module	Type of Error	Error_Messages
2016-01-01T00:00:00.000+05:30	1	1	2222	Error	Hadoopcluster1	p121	t123	module1	error_type-1	error-message-1
2016-01-01T00:00:00.000+05:30	1	1	2224	Error	Hadoopcluster3	p123	t127	module1	error_type-2	error-message-2
2016-01-01T00:00:00.000+05:30	1	1	2223	Error	Hadoopcluster2	p122	t125	module1	error_type-3	error-message-3
2016-01-01T00:02:00.000+05:30	1	1	2222	Error	Hadoopcluster1	p121	t123	module1	error_type-3	error-message-4
2016-01-01T00:02:00.000+05:30	1	1	2224	Error	Hadoopcluster3	p123	t127	module1	error_type-2	error-message-5

THE SCRIPT THAT WILL GENERATE THESE THREE TABLES IS

NEXT WE HAVE TO GENERATE
ERROR_SUMMARY

```
grunt> log_data = load 'log_data.txt';
grunt> log_data = filter log_data by not $3 == 'message';
grunt> log_data = foreach log_data generate SUBSTRING($0,1,10),SUBSTRING($0,11,10),SUBSTRING($0,12,10),SUBSTRING($0,13,10),SUBSTRING($0,14,10),SUBSTRING($0,15,10),SUBSTRING($0,16,10),SUBSTRING($0,17,10),SUBSTRING($0,18,10),SUBSTRING($0,19,10);
grunt> log_data = foreach log_data generate CONCAT($0,$1,$2,$3,$4,$5,$6,$7,$8,$9,$10);
grunt> log_data = foreach log_data generate CONCAT($0,$1),$2,$3,$4,$5,$6,$7,$8,$9;
grunt> log_data = foreach log_data generate ToDate($0,'yyyy-mm-dd HH:mm:ss'),$1,$2,$3,$4,$5,$6,$7,$8;
grunt> log_data = foreach log_data generate $0 as Time_Stamp,GetMonth($0) as Month,GetDay($0) as Day,$1 as server_number,$2 as type_of_message,$3 as cluster_name,$4 as process_ID,$5 as thread_ID,$6 as module,$7 as error_type,$8 as error_message;

grunt> split log_data into warnings_data if type_of_message == 'warning', error_data if type_of_message == 'Error';
grunt> groupd = group error_data by (server_number,error_type);
```

```
grunt> error_summary = foreach groupd{
grunt>   number_of_occurrences = COUNT(error_data);
grunt>   error_message = error_data[0].error_message;
grunt>   error_type = error_data[0].error_type;
grunt>   sort error_data by error_type;
grunt>   top 1 in error_data by error_type;
grunt>   generate error_summary from error_data;
grunt>   groupd = groupd - error_data;
grunt> };

```

Server_Number	Error_Type	Last_Time_Stamp	Process ID	Thread ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module1	1	error-message-4

```
grunt> store error_summary into 'error_summary';
grunt> store error_data into 'error_data';
grunt> store warnings_data into 'warnings_data';
```

FIRST WE DO A GROUP BY ON
server_number,error_type

FIRST WE DO A GROUP BY ON
server_number, error_type

**THIS PART OF THE
SCRIPT WILL
GENERATE THIS TABLE**

```
grunt> error_summary = foreach group{
  grunt>   number_of_occurrences = COUNT(error_data);
  grunt>   error_messages = BagToString(error_data.error_message,',');
  grunt>   error_data = foreach error_data generate server_number, error_type,
    Time_Stamp, process_ID, thread_ID, module;
  grunt>   sorted_error_messages = order error_data by Time_Stamp desc;
  grunt>   top_item = limit sorted_error_messages 1;
  grunt>   generate flatten(top_item),number_of_occurrences,error_messages;
  grunt> };

  grunt> store error_summary into 'error_summary';
  grunt> store error_data into 'error_data';
  grunt> store warnings data into 'warnings data';
```

FIRST WE DO A GROUP BY ON server_number,error_type

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module1	1	error-message-4

THIS LINE USES
AGGREGATE FUNCTION TO
COUNT THE FREQUENCY OF
THE ERRORS

```
grunt> error_summary = foreach groupd{  
grunt>   number_of_occurrences = COUNT(error_data);  
grunt>   error_messages = BagToString(error_data.error_message,',');  
grunt>   error_data = foreach error_data generate server_number, error_type,  
Time_Stamp, process_ID, thread_ID, module;  
grunt>   sorted_error_messages = order error_data by Time_Stamp desc;  
grunt>   top_item = limit sorted_error_messages 1;  
grunt>   generate flatten(top_item),number_of_occurrences,error_messages;  
grunt> };  
  
grunt> store error_summary into 'error_summary';  
grunt> store error_data into 'error_data';  
grunt> store warnings_data into 'warnings_data';
```

FIRST WE DO A GROUP BY ON

server number, error type

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module1	1	error-message-4

THIS LINE USES
AGGREGATE FUNCTION TO
AGGREGATE ALL THE
MESSAGES
CORRESPONDING TO ONE
ERROR MESSAGE

```
grunt> error_summary = foreach groupd{  
grunt>   number_of_occurrences = COUNT(error_data);  
grunt>   error_messages = BagToString(error_data.error_message,',');  
grunt>   error_data = foreach error_data generate server_number, error_type,  
Time_Stamp, process_ID, thread_ID, module;  
grunt>   sorted_error_messages = order error_data by Time_Stamp desc;  
grunt>   top_item = limit sorted_error_messages 1;  
grunt>   generate flatten(top_item),number_of_occurrences,error_messages;  
grunt> };  
  
grunt> store error_summary into 'error_summary';  
grunt> store error_data into 'error_data';  
grunt> store warnings data into 'warnings data';
```

FIRST WE DO A GROUP BY ON server_number, error_type

```
grunt> log_data = load 'log_data.txt' using PigStorage('');
```

```
grunt> log_data = filter log_data by not $3 == 'message';
```

```
grunt> log_data = foreach log_data generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;
```

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module1	1	error-message-4

WE NEED ONLY THESE
FIELDS IN THE FINAL
OUTPUT

```
grunt> error_summary = foreach groupd{
```

```
grunt> number_of_occurrences = COUNT(error_data);
```

```
grunt> error_messages = BagToString(error_data.error_message,',');
```

```
grunt> error_data = foreach error_data generate server_number, error_type,
```

```
Time_Stamp, process_ID, thread_ID, module;
```

```
grunt> sorted_error_messages = order error_data by Time_Stamp desc;
```

```
grunt> top_item = limit sorted_error_messages 1;
```

```
grunt> generate flatten(top_item),number_of_occurrences,error_messages;
```

```
grunt> };
```



```
grunt> store error_summary into 'error_summary';
```

```
grunt> store error_data into 'error_data';
```

```
grunt> store warnings_data into 'warnings_data';
```

FIRST WE DO A GROUP BY ON server_number, error_type

```
grunt> log_data = load log_data.txt using PigStorage(',');
grunt> log_data = filter log_data by not $3 == 'message';
grunt> log_data = foreach log_data generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;
```

```
grunt> log_data = foreach log_data generate NCAT($0,' ') $2 $3 $4 $5 $6 $7,$8,$9,$10;
      ,$8,$9;
      ),$1,$2,$3,$4,$5,$6,$7,$8;
      s Month,GetDay($0,$1) server_id,$2 as
      module,$7 as error_type,GetHour($0,$1) as
      hour_data if type(error_type) != 'string' then
      'error'
```

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module1	1	error-message-4

```
grunt> error_summary = foreach groupd{
```

```
grunt> number_of_occurrences = COUNT(error_data);
```

```
grunt> error_messages = BagToString(error_data.error_message,',');
```

```
grunt> error_data = foreach error_data generate server_number, error_type,
```

```
      $1 as server_id, $2 as thread_id, $3 as
```

```
grunt> sorted_error_messages = order error_data by Time_Stamp desc;
```

```
grunt> top_item = limit sorted_error_messages 1;
```

```
grunt> generate flatten(top_item),number_of_occurrences,error_messages;
```

```
grunt> };
```

```
grunt> store error_summary into 'error_summary';
```

```
grunt> store error_data into 'error_data';
```

```
grunt> store warnings_data into 'warnings_data';
```

**WE ORDER THIS RELATION
ON TIME STAMP TO
FIND THE LATEST TIME
STAMP WHEN AN ERROR
OCCURRED**

FIRST WE DO A GROUP BY ON server_number,error_type

```
grunt> log_data = load log_data.txt using PigStorage(',');
grunt> log_data = filter log_data by not $3 == 'message';
grunt> log_data = foreach log_data generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;
```

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module1	1	error-message-4

WE LIMIT THIS RELATION
TO ONLY ONE ROW

```
grunt> error_summary = foreach groupd{
  number_of_occurrences = COUNT(error_data);
  error_messages = BagToString(error_data.error_message,',');
  error_data = foreach error_data generate server_number, error_type,
  Time_Stamp, process_ID, thread_ID, module;
  sorted_error_messages = order error_data by Time_Stamp desc;
grunt> top_item = limit sorted_error_messages 1;
  generate flatten(top_item),number_of_occurrences,error_messages;
  };
  store error_summary into 'error_summary';
  store error_data into 'error_data';
  store warnings_data into 'warnings_data';
```

FIRST WE DO A GROUP BY ON server_number,error_type

```
grunt> log_data = load('log_data.txt') using PigStorage(',')  
grunt> log_data = filter log_data by not $3 == 'message';  
grunt> log_data = foreach log_data generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;  
grunt> log_data = foreach log_data generate SERVER_NUMBER($0),NCAT($0,' ') ,$2,$3,$4,$5,$6,$7,$8,$9,$10;  
,$8,$9;  
,,$1,$2,$3,$4,$5,$6,$7,$8;  
is Month,GetDay($0) as Day,$1 as server_number,$2 as  
module,$7 as error_type ,,$8 as error_message;  
rror_data = foreach log_data generate SERVER_NUMBER($0) as server_number,  
$1 as module,$2 as error_type ,,$3 as error_message;
```

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module1	1	error-message-4

```
grunt> error_summary = foreach groupd{  
grunt> number_of_occurrences = COUNT(error_data);  
grunt> error_messages = BagToString(error_data.error_message,',');  
grunt> error_data = foreach error_data generate server_number, error_type,  
Time_Stamp, process_ID, thread_ID, module;  
grunt> sorted_error_messages = order error_data by Time_Stamp desc;  
grunt> top_item = limit sorted_error_messages 1;  
grunt> generate flatten(top_item),number_of_occurrences,error_messages;  
grunt> };
```

```
grunt> store error_summary into 'error_summary';  
grunt> store error_data into 'error_data';  
grunt> store warnings_data into 'warnings_data';
```

THIS COMMAND FLATTENS

top_item AND
GENERATES THESE
COLUMNS

FIRST WE DO A GROUP BY ON

server_number, error_type

```
grunt> log_data = load 'log_data.txt';
grunt> log_data = filter log_data by not $3 == 'message';
grunt> log_data = foreach log_data generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;
```

Server_Number	Error_Type	Last_Time_Stamp	Process_ID	Thread_ID	Module	Number_of_occurrences	Error_Messages
2222	error_type-1	2016-01-01T00:00:00.000+05:30	p121	t123	module1	1	error-message-1
2222	error_type-3	2016-01-01T00:02:00.000+05:30	p121	t123	module1	1	error-message-4

FOR EVERY SERVER
NUMBER AND ERROR TYPE

```
grunt> error_summary = foreach groupd{
```

```
grunt> number_of_occurrences = COUNT(error_data);
```

WE'VE GOT THE LAST TIMESTAMP OF THE ERROR AND AN
AGGREGATED NUMBER OF ERRORS OF THAT TYPE ON THAT SERVER

```
grunt> sorted_error_messages = order error_data by Time_Stamp desc;
```

```
grunt> top_item = limit sorted_error_messages 1;
```

```
grunt> generate flatten(top_item),number_of_occurrences,error_messages;
```

```
grunt> };
```

```
grunt> store error_summary into 'error_summary';
```

```
grunt> store error_data into 'error_data';
```

```
grunt> store warnings_data into 'warnings_data';
```

FIRST WE DO A GROUP BY ON

server_number, error_type

```
grunt> log_data = load('logdata') using LogLoader;
grunt> log_data = filter(log_data, not(isEmpty(log_data)));
grunt> log_data = foreach log_data generate SUBSTRING($0,1,11),$1,SUBSTRING($2,0,8),$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,' '),$2,$3,$4,$5,$6,$7,$8,$9,$10;
grunt> log_data = foreach log_data generate CONCAT($0,$1),$2,$3,$4,$5,$6,$7,$8,$9;
grunt> log_data = foreach log_data generate ToDate($0,'yyyy-mm-dd HH:mm:ss'),$1,$2,$3,$4,$5,$6,$7,$8;
grunt> log_data = foreach log_data generate $0 as Time_Stamp,GetMonth($0) as Month,GetDay($0) as Day,$1 as server_number,$2 as type_of_message,$3 as cluster_name,$4 as process_ID,$5 as thread_ID,$6 as module,$7 as error_type,$8 as error_message;

grunt> split log_data into warnings_data if type_of_message == 'warning', error_data if type_of_message == 'Error';
grunt> groupd = group error_data by (server_number,error_type);
```

**STORES THE DATA OF
VARIOUS RELATIONS INTO
THE HDFS DIRECTORY**

```
grunt> error_summary = foreach groupd{
grunt>   number_of_occurrences = COUNT(error_data);
grunt>   error_messages = BagToString(error_data.error_message, ',');
grunt>   error_data = foreach error_data generate server_number, error_type,
Time_Stamp, process_ID, thread_ID, module;
grunt>   sorted_error_messages = order error_data by Time_Stamp desc;
grunt>   top_item = limit sorted_error_messages 1;
grunt>   generate flatten(top_item),number_of_occurrences,error_messages;
grunt> };

grunt> store error_summary into 'error_summary';
grunt> store error_data into 'error_data';
grunt> store warnings_data into 'warnings_data';
```