Analysis of Afghan War Dairies

Use Case

Version 1.0

Analysis of Afghan War Dairies

The data were written by soldiers and intelligence officers of the United States Military. Too keep it simple, we will analyze only four of the available columns (Type, Category, Region and Attack On) in the dataset. The dataset is made available in next section.

Problem Statements:

Below are few of the problem statement that we have chosen to work on this dataset.

In Pig

- To examine all events that involve explosive hazards.
- To examine explosive events that involve Improvised Explosive Devices (IEDs).

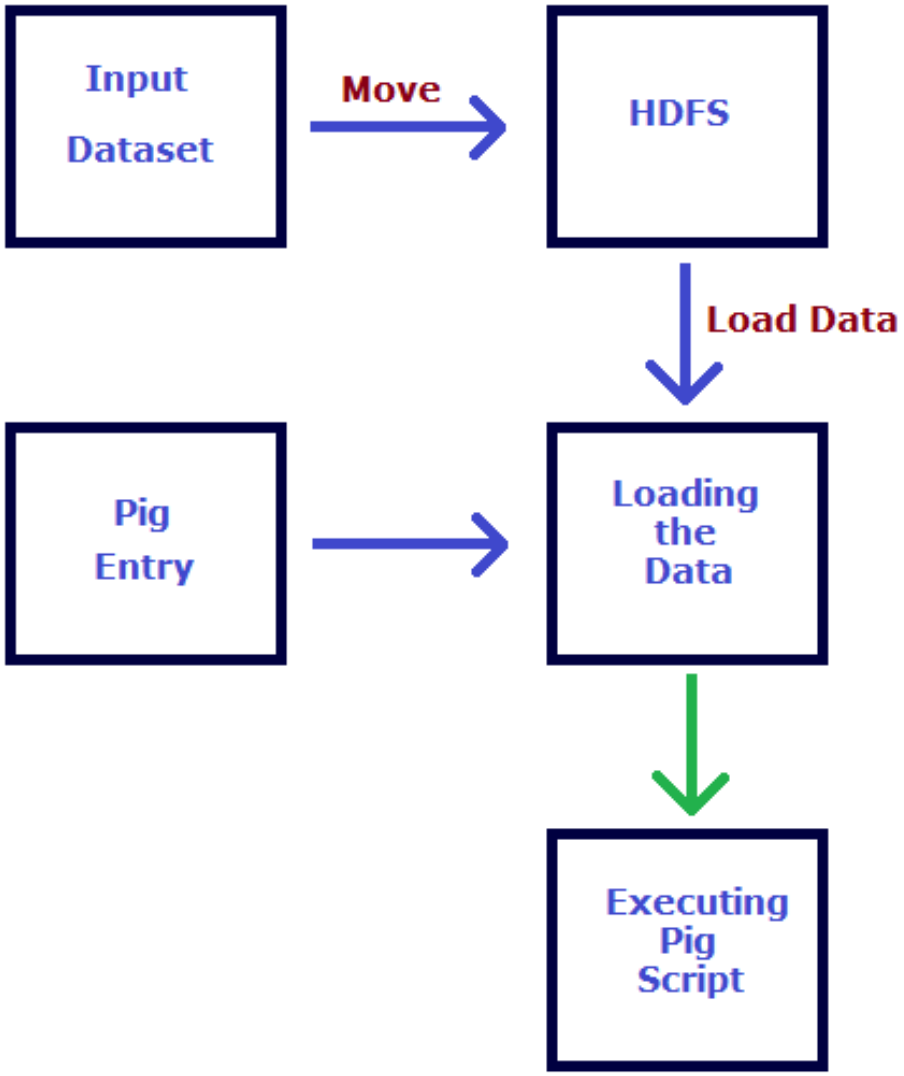
Important Links:

- **Complete Dataset:**
  https://www.google.com/fusiontables/DataSource?dsrcid=224453#rows:id=1
- **Used Dataset:**
  https://edureka.wistia.com/medias/gv22zpwisk/download
- **Link for all the codes:**
  https://edureka.wistia.com/medias/r7egjwnrdn/download?media_file_id=66604812
- **Pig Installation:**
  https://edureka.wistia.com/medias/lpb6yiupps

Technology/Software Used:

- Hadoop environment
- Apache Pig

Project Workflow:



Environment Creation for Creating Solution

To execute the problem statements, first we have downloaded the data file in form of CSV file (link is on page 1). For simplicity, we have renamed the file and moved the file to HDFS. For moving the file from local to HDFS, command is:

hadoop dfs -copyFromLocal /home/edureka/afghan.csv hdfs:/

Pig Problem 1:

To examine all events that involve explosive hazards:

Here, we have used pig script to get the output which is shown below:

```
File  Edit  View  Search  Terminal  Help
[edureka@localhost Downloads]$ pig solution1.pig
2015-01-10 05:11:08,981 [main] INFO  org.apache.pig.Main - Apache Pig version 0.12.
2015-01-10 05                                                              sages t
2015-01-10 05                                                              - Unab
n-java classe
2015-01-10 05                                                              t bootu
2015-01-10 05                                                              engine.
alhost:8020
2015-01-10 05                                                              te - Pi
2015-01-10 05                                                              er.Logic
uplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, Li
PartitionFilterOptimizer, PartitionFilterOptimizer, PushDownForEachFlatten, PushUpF
```

```
A = LOAD '/afghan.csv' USING PigStorage(',') AS (type:chararray,
category:chararray, region:chararray, attack:chararray);
B = filter A by type == 'Explosive Hazard';
C = group B by region;
D = foreach C generate FLATTEN(group) as region, COUNT(B) as total;
E = order D by total ASC;
DUMP E;
```

```
2015-01-10 05:12:37,345 [main] INFO  org.apac
(NONE SELECTED,4)
(UNKNOWN,82)
(RC CAPITAL,746)
(RC NORTH,844)
(RC WEST,1009)
(RC EAST,8989)
(RC SOUTH,11408)
[edureka@localhost Downloads]$
```

Output

Pig Problem 2:

To examine explosive events that involve Improvised Explosive Devices (IEDs):

```
File  Edit  View  Search  Terminal  Help
[edureka@localhost Downloads]$ pig solution2.pig
2015-01-10 05:11:08,981 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.
2015-01-10 05                                                              sages t
2015-01-10 05                                                              - Unab
n-java classe
2015-01-10 05                                                              t bootu
2015-01-10 05                                                              engine.
alhost:8020
2015-01-10 05                                                              te - Pi
2015-01-10 05:                                                             er.Logic
uplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, Li
PartitionFilterOptimizer, PartitionFilterOptimizer, PushDownForEachFlatten, PushUpF
```

```
A = LOAD '/afghan.csv' USING PigStorage(',') AS (type:chararray,
category:chararray, region:chararray, attack:chararray);
B = filter A by category == 'IED Explosion';
C = filter A by category == 'IED Found/Cleared';
D = cogroup B by category , C by category;
E = foreach D generate COUNT(B), COUNT(C);
DUMP E;
```

```
2015-01-10 05:26:00,214 [main] INFO
2015-01-10 05:26:00,214 [main] INFO
(7202,0)
(0,8581)
[edureka@localhost Downloads]$
```

Output