|  |
| --- |
|  |
|  | **Problem Statement 1:**    Find out the top 5 most visited destinations.  **Solution:** |
|  |  |
|  | -- Register piggybank.jar to use CSVExcelStorage |
|  | REGISTER '/home/acadgild/ankita/Downloads/piggybank.jar'; |
|  |  |
|  | --Load flight details |
|  | flight\_data = load '/home/acadgild/ankita/Assignment5\_2/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER'); |
|  |  |
|  | -- Get the destinations |
|  | flight\_dest\_details = foreach flight\_data generate (chararray) $18 as dest; |
|  |  |
|  | -- Filter the destinations - get only not null destinations |
|  | dest\_data = filter flight\_dest\_details by dest is not null; |
|  |  |
|  | -- group result by destination |
|  | dest\_group = group dest\_data by dest; |
|  |  |
|  | -- count destinations for every group |
|  | dest\_count = foreach dest\_group generate group, COUNT(dest\_data.dest); |
|  |  |
|  | -- sort results by count of destinations in descending order |
|  | top\_dest = order dest\_count by $1 DESC; |
|  |  |
|  | -- fetch top 5 rows from the result |
|  | top5\_dest = LIMIT top\_dest 5; |
|  |  |
|  | -- load Airport details |
|  | airport\_data = load '/home/acadgild/ankita/Assignment5\_2/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER'); |
|  |  |
|  | -- get destination, city and country |
|  | airport\_country\_city = foreach airport\_data generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country; |
|  |  |
|  | -- join flight data and airport data to get destination's country and city |
|  | dest\_country\_city = join top5\_dest by $0, airport\_country\_city by dest; |
|  |  |
|  | -- sort final result by count of destinations |
|  | final\_result = order dest\_country\_city by $1 DESC ; |
|  |  |
|  | -- display the results |
|  | dump final\_result;  **Output:**     |  | | --- | |  | |  | **Problem Statement 2 :**  Which month has seen the most number of cancellations due to bad weather?  **Solution:** | |  |  | |  | -- Register piggybank.jar to use CSVExcelStorage | |  | REGISTER '/home/acadgild/ankita/Downloads/piggybank.jar'; | |  |  | |  | --Load flight details | |  | flight\_data = load '/home/acadgild/ankita/Assignment5\_2/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER'); | |  |  | |  | -- Get the month, cancelled and cancellation code | |  | flight\_details = foreach flight\_data generate (chararray) $2 as month,(int) $22 as cancelled, (chararray)$23 as reason; | |  |  | |  | -- Filter the flight details - get only cancelled flights due to bad weather | |  | cancelled\_flights\_badweather = filter flight\_details by cancelled==1 AND reason=='B'; | |  |  | |  | -- group results by month | |  | group\_cancelled\_flights\_by\_month = group cancelled\_flights\_badweather by month; | |  |  | |  | -- count bad weather cancellations by month | |  | count\_cancelled\_flights\_by\_month = foreach group\_cancelled\_flights\_by\_month generate group, COUNT(cancelled\_flights\_badweather.reason); | |  |  | |  | -- sort results by count of cancellations in descending order | |  | sort\_cancellation\_by\_month = order count\_cancelled\_flights\_by\_month by $1 DESC; | |  |  | |  | -- fetch top row from the result | |  | top\_cancellation\_month = LIMIT sort\_cancellation\_by\_month 1; | |  |  | |  | dump top\_cancellation\_month ;  **Output:** |      |  | | --- | |  | |  | **Problem Statement 3:**  Top ten origins with the highest AVG departure delay  **Solution:** | |  |  | |  |  | |  | -- Register piggybank.jar to use CSVExcelStorage | |  | REGISTER '/home/acadgild/ankita/Downloads/piggybank.jar'; | |  |  | |  | --Load flight details | |  | flight\_data = load '/home/acadgild/ankita/Assignment5\_2/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER'); | |  |  | |  | -- Get the month, cancelled and cancellation code | |  | flight\_details = foreach flight\_data generate (int) $16 as delay,(chararray) $17 as origin; | |  |  | |  | -- Filter the flight details - delay time should not be null and origin should not be null | |  | flight\_details\_filter = FILTER flight\_details by origin is not null and delay is not null ; | |  |  | |  | -- group result by origin | |  | group\_by\_origin = group flight\_details\_filter by origin; | |  |  | |  | -- get the average of delays by origin | |  | sum\_delay = foreach group\_by\_origin generate group, AVG(flight\_details\_filter.delay); | |  |  | |  | -- sort results by descending order of delay | |  | sort\_flight\_delays = order sum\_delay by $1 DESC; | |  |  | |  | -- fetch top 10 rows from the result | |  | top10\_flight\_delays = LIMIT sort\_flight\_delays 10; | |  |  | |  |  | |  | -- load Airport details | |  | airport\_data = load '/home/acadgild/ankita/Assignment5\_2/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER'); | |  |  | |  | -- get destination, city and country | |  | airport\_country\_city = foreach airport\_data generate (chararray)$0 as origin, (chararray)$2 as city, (chararray)$4 as country; | |  |  | |  | -- join flight data and airport data to get destination's country and city | |  | dest\_country\_city = join top10\_flight\_delays by $0, airport\_country\_city by origin; | |  |  | |  | -- sort final result by count of destinations | |  | final\_result = order dest\_country\_city by $1 DESC ; | |  |  | |  | -- display the results | |  | dump final\_result;  **Output:** |  |  | | --- | |  | |  | |  | **Problem Statement 4:**  Which route (origin & destination) has seen the maximum diversion?  **Solution:** | |  |  | |  |  | |  | -- Register piggybank.jar to use CSVExcelStorage | |  | REGISTER '/home/acadgild/ankita/Downloads/piggybank.jar'; | |  |  | |  | --Load flight details | |  | flight\_data = load '/home/acadgild/ankita/Assignment5\_2/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER'); | |  |  | |  | -- Get the origin, destination and diverted details | |  | flight\_details = foreach flight\_data generate (chararray) $17 as origin,(chararray)$18 as dest, (int)$24 as diverted; | |  |  | |  | -- Filter the flight details - origin and destination should not be null and it should be diverted. | |  | flight\_details\_filter = FILTER flight\_details by origin is not null and dest is not null and diverted==1 ; | |  |  | |  | -- group result by origin and destination | |  | group\_by\_origindest = group flight\_details\_filter by (origin,dest); | |  |  | |  | -- get the number of diversions | |  | count\_diversions = foreach group\_by\_origindest generate group, COUNT(flight\_details\_filter.diverted); | |  |  | |  | -- sort results by descending order of diversions | |  | sort\_flight\_diversions = order count\_diversions by $1 DESC; | |  |  | |  | -- fetch top 10 rows from the result | |  | top10\_flight\_diversions = LIMIT sort\_flight\_diversions 10; | |  |  | |  | dump top10\_flight\_diversions ;  **Output:** | |  |  | |