**The Client's Question (Business problem)** - How much revenue is the client getting from external Search Engines, such as Google, Yahoo and MSN, and which keywords are performing the best based on revenue?

**Dataset:** Hit Level Data

IP is the unique key which tells all the different activities that a user does while searching for a keyword till he completes the purchase. Here is the sequence of actions done by the user,

* Search for the product using any of the search engine (event list:2)
* Add the product to the shopping cart (event list:12)
* Fill the order checkout details in shopping cart (event list:11)
* Get an order confirmation (event list:NaN)
* Purchase order completion (event list:1)

Referrer URL has the search engine which is the hostname and search keyword details from the query details from parsing the URL.

Once the order is complete, product list column has the actual revenue gained out of the order

Both search details and revenue can be joined together using the common IP key

On sorting the joined results based on revenue with descending, we can answer which search engine generated more revenue on which product

|  | **Search Engine Domain** | **Search Keyboard** | **Revenue** |
| --- | --- | --- | --- |
|  | www.google.com | Ipod | 290 |
|  | www.bing.com | Zune | 250 |
|  | www.google.com | ipod | 190 |

**Github Repo:** <https://github.com/aruntr2000/data_engineering>

**Implementation details:**

Local implementation of the business problem with filename as argument to the class

<https://github.com/aruntr2000/data_engineering/blob/master/data_processing_local_run.py>

Serverless Deployment of the business problem in AWS:

Python script creates S3 bucket and Data Processing Lambda in AWS. It uploads the data file into S3 bucket input prefix, which in turn triggers the Data Processing Lambda to process and create output file into S3 bucket output prefix

<https://github.com/aruntr2000/data_engineering/blob/master/setup_app.py>

<https://github.com/aruntr2000/data_engineering/blob/master/data_processing_lambda.py>

Python scripts deletes the S3 bucket and Data Processing Lambda in AWS when it is not needed <https://github.com/aruntr2000/data_engineering/blob/master/cleanup_app.py>

Diagram

Description automatically generated with medium confidence