# **Data Platform Engineer**

## Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Part 1 – Technical Test

You have a directory with files in the following format:

**video\_id,ip,event\_type**

Column event\_type can be any value of: **view, click, impression**

A sample file will look like this:

*1, 64.236.4.133, view*

*1, 64.236.4.133, view*

*2, 64.236.4.133, impression*

*2, 54.72.128.142, click*

Please write a Map Reduce Java program that will count how many views, impressions or clicks there are in the file per video\_id and country\_code. To detect the country code you will need to use the geoip library. Easiest way to use it is by adding a maven dependency from search.maven.org.

Format of the output file is:

**video\_id, country\_code, views, impressions, clicks**

Output for the above input will look like this:

*1, UK, 2, 0, 0*

*2, UK, 0, 1, 0*

*2, US, 0, 0, 1*

**Bonus** points if the output is produced as avro files or written directly to a Hive table.

**Expectations**

Write the **best code** you can. Demonstrate you design and automated testing skills. It's **important** for us to see you can write quality code and **not** just code that works. Think about interfaces, abstractions, testability, performance. If possible include a maven project file pom.xml and instructions with your program to make it easier for us to run it.

**Timelines**

Please complete the assignment within 7 days. You will receive **bonus** points if you complete it in 3 days.

## Part 2 – Bonus Home Assignment (Optional)

**Syndication Platform Data Warehouse (DWH)**

Imagine a platform that allows content owners (creator of videos) to publish their videos on many web sites with lots of traffic and earn money from ads played on their videos.

The platform should handle a massive amount (1 TB per day) of log files from different data sources. Part of the logs come from the video player with a file for each event on the video. For example *VideoStarted, VideoEnded* events would come in separate log files and have different parameters. Video files come in as a stream and can be uploaded to you continuously.

Other part of the logs come from the advertisement videos from a different third party system. These logs are made available to you every hour on an FTP server.

Business wants to have near real time reports (15 min delay) on how well videos and ads are performing.

**Sample Reports**

1. Performance report - how many video views and ad views we had on each web site filtered by date, country, device, video, viewer gender, content owner
2. Revenue report - describes how much revenue each video has earned filtered by date, web site, country, device, gender, content owner

**Requirements:**

1. Business analysts should be able to run different kinds of unknown queries on the DWH
2. Other third party data warehouses need to be able to receive new data from the DWH on a daily basis.
3. In general the platform should be very stable and performant even in case of hardware failures.

**Question 1**

Propose a general architecture of the DWH that will handle these requirements. Draw a diagram that will describe each part in the system. Please explain your choice of technologies for each part.

**Question 2**

Explain in more detail the how the platform will receive data from different data sources before being processed.

**Question 3**

Explain what will be the data flow inside your DWH. How the data will be stored in each part of the system.

**Question 4**

Explain how different users will access the data in your DWH.