**The Story**

Imagine that there is a stream which ingests 300 million raw data in a day.

• Data comes from various sources. e.g. web sites, mobile apps.

• It consists of users’ clickstream and events data according to their behaviors.

• Each user has a unique session id. Sessions expire in 30 minutes if the user is not on the web site/mobile app up to 30 minutes.

• In special days (e.g Black Friday, Christmas Week) delivered amount of data can be higher than normal day. It can reach 500 million raw data in a day.

**Needs**

• Keeping the system scalable, low latency and fault-tolerant.

• Analytics of the data to be able to report all the sessions/users.

• Processing the data in real-time. e.g Last visits, purchases, cart amounts, pre-defined events (add to cart, whishlisted products) of each user.

• Making predictions of the data in daily basis.

• Recognize users across different devices. To be able to make unficiation.

**Requirements**

1- Draw a lambda architecture considering all the needs and it should operate on huge amount of data as mentioned above.

• Specify the technologies that you will use in all the layers.

• Explain why you use the technologies that you will design.

2- Choose one of the layers of the architecture that you draw above and develop it in a small-scale.

• Use sample dataset below which data is coming from multiple sources. Transform the data and extract some features which will be used to predict **likelihood to purchase** of a user.

• You should use at least one NoSQL database to persist them. e.g. Cassandra, HBase.

• You can use a data processing engine. e.g Spark, Storm.

• You can use a messaging system. e.g. Kafka, Kinesis.

**Sample data**

https://drive.google.com/open?id=1Fvp9HaxBIvJqL16z2od1taAyjTtrs1tL

Example event:

{

"session\_id": "995a888d-69a9-3779-9c91-0bb7ff7114c8\_1510761641",

"event": "pageView",

"partner\_id": "00215",

"partner\_name": "lcwaikiki",

"cart\_amount": 0,

"country": "TR",

"user\_agent": "Mozilla/5.0 (Linux; Android 6.0.1; SAMSUNG SM-J700F Build/MMB29K) AppleWebKit/537.36 (KHTML, like Gecko) SamsungBrowser/5.4 Chrome/51.0.2704.106 Mobile Safari/537.36",

"user\_id": "3939a171d2ce3.4126983939a",

"version": "1.0",

"language": "tr\_TR",

"date": 1510758058,

"search\_query": "tisort-body-ve-atlet",

"current\_url": "http://www.lcwaikiki.com/tr-TR/TR/kategori/kiz-bebek/tisort-body-ve-atlet",

"category": ["Kız Bebek", "Tişört, Body ve Atlet"],

"referrer": "http://www.lcwaikiki.com/tr-TR/TR",

"init\_session": false,

"page\_type": "category"

}

• “partner\_name” can be changed. (e.g. 10 different partners)

• “user\_id” is unique for each client.

• “session\_id” is unique for each user and it expires in 30 minutes in client side.

• “page\_type” can take 6 values. ("main", "product", "category", "cart", "success", "other")

• “init\_session” field becomes true/false and creates “session” event if the user opens a new session.

* **After you complete the test please send the result to parivallal.radhakrishnan@moneysmart.com with subject:** [MS Data Engineer] Application by “Name Surname”