# Technical Lead: Data Ingestion Interview

## Objective

Imagine that Alaska Airlines has just finished developing their brand-new E-Commerce site. Unfortunately, their development team was unaware of the need for data tagging. They have hired you to come help them define their strategy for tagging key elements of their site. For this exercise, we would like you to focus specifically on modeling customer interactions around the flight search space.

In this exercise, we will ask you to exemplify how you would model, capture, and store data around the flight search space. Ultimately, we will want the data to, *at a minimum*, be query-able to answer the following questions:

1. What is the conversion rate for an individual flight search result? For definition, an ‘individual flight search result’ is the following:

Conversion rate pertaining to this particular scenario would be number of users visiting this page against the number of users that selected one of the fair types and progressed to next pages of flight booking.

Data needed to capture from a storage/retrieval perspective:  
- Number of users visiting this page/Number of users exiting this page without selection  
- No of users selected at least one fare type  
- Dates, Source and destination cities and type of flight

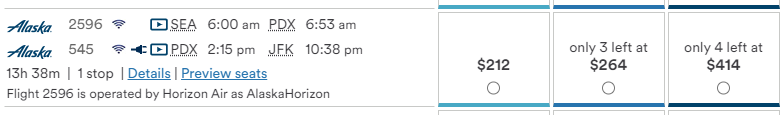
Data for tagging/analytics perspective (limited to below scenario):

* User IP
* Session ID
* Region of access
* User details for existing member (PII or otherwise)
  + Email, address, Ph number
* Browser
* Cookies

The incoming data would need to be modelled into following tables

* Flight details
* Airport details
* Fare

However, the eventual profitable conversion rate for any transaction of visiting customer would be in making business and booking a ticket.



1. How does the cost in miles affect the guests’ willingness to purchase? For this question, assume that the conversion space is already tagged successfully, and this data is already accessible.

(Dependency: I cannot determine what’s profitable to business – 414$ or a 23k(for example) miles for business class)

Assumptions:

1. Conversion space here is user opting miles over dollar payment:
2. Enrolling in miles program when the option seems profitable:

Data needed to be captured:

* Similar to above scenario.

Data needed to capture from a storage/retrieval perspective:

* Cost in miles for this to/from distance on avg cost

The incoming data would need to be modelled into following tables

* Flight details
* Airport details
* Fare
* Member

Here comes the *profitable* conversion rate in terms of revenue based on dependency above.

Metrics on whether a user has selected either of assumed goals would answer the query.

* For an existing user, if we could achieve the next level fair type selection through cost in miles/flyer programs – yes
* Even if it’s a guest user, we can suggest and average cost per miles would cost for that distance (Involves back end logic) he’s travelling. If less, we get a new member.

Note:

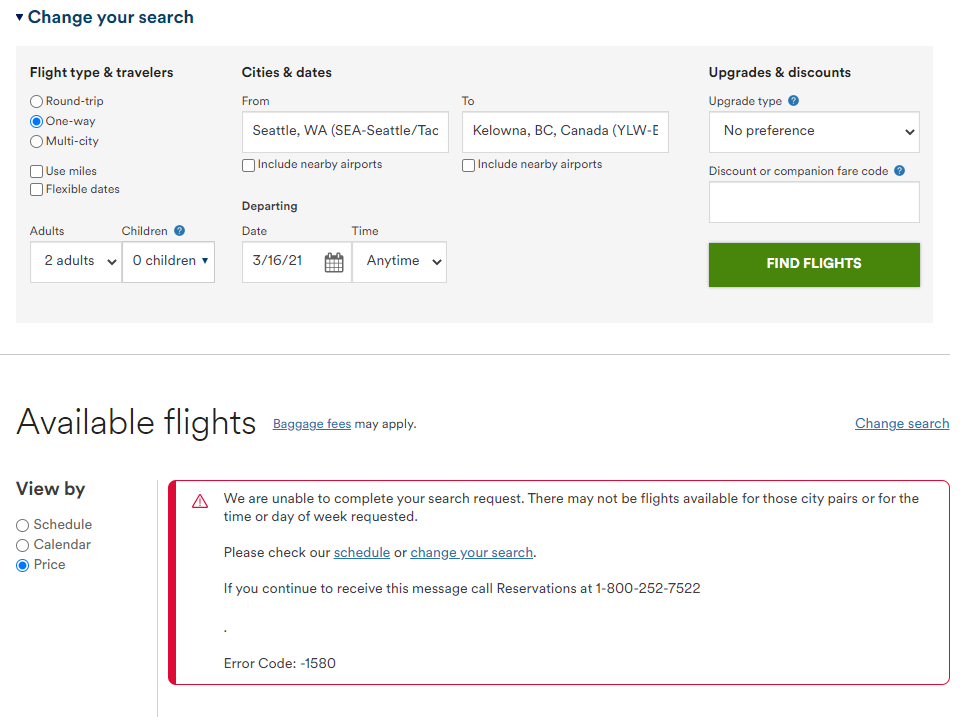
Involves overhead of multiple calls to back end.  
Performance Impact on backend calculations

1. How many times does a guest attempt to search for a destination that we do not serve for the given airport or date? This can be demoed through the following flight search

Clarification needed: What would be the impact on multiple search?

Data on session/input parameters (airport, dates) as well as call made would help cache the call preventing unnecessary calls to back end.

Data needed for tracking/storage is pretty much similar to above calls.



## Tasks

1. Head over to alaskaair.com and perform a search for any flight. Consider the different elements of data that could be captured through this interaction. We would recommend you *at least* focus on Flight Search and Flight Results as possible events to model, but feel free to explore beyond those two if you feel further structure is necessary. Modeling these events as a JSON Schema is preferred.

In addition to details discussed on a higher level, I would like to go over the modelling part of it in details here.

Out of several event modeling techniques, I would choose traditional common event table and Json to retrieve general attributes purely based on following reasons:

* Ease of use and less complexity 😊
* Easier analysis on common attributes such as session and user ids
* BI tools can access underlying base storage tables to get around limitations in json access

Json request schema sample:

{

Session: [{

id: 324572,

IP: 192.168 .1 .1,

Country: USA,

timestamp: 16679042407,

Device: Android,

User: [{

Id: 658360232,

Email: xcv @hotmail.com,

timestamp: 16679042407,

Phone: +1617343 XXXX

}]

Flight: [{

Number: AA937,

Source: RDU,

Destination: BOS,

timestamp: 16679042407

}]

}]

Each payload is tied back to corresponding storage table on internal logic so as to load respective attributes into following tables.

* Session
* Member
* Flight

Concerns:

Performance optimization between joining tables

1. Describe the tools/process you would use to capture these events as they occur. Feel free to include 3rd party tools as part of this process.

Tools:

* Google Analytics with GTM
* Tealium extensions
* Hotjar

1. Propose at least two different options of performing this data capture and list each method’s pros and cons. You are welcome to include more than two architectures of tagging if you believe there is reason to do so.

1. The capture strategy will pipe directly to a storage strategy. Based on how you chose to model data and how you chose to capture it, please walkthrough the structure the data would have in a database. Please provide at least two methodologies here, listing their pros and cons.

## Process

Create an open source repository in Github to store any information related to your solution. If you would like to give us an overview of how the solution should be read through, please write a Solution.md file to walk us through how we should examine your solution.

You are **not required** to develop, write pipelines, or create a local database as part of this project, but please be verbose in your explanations. Good communication will be key to success in this role.

## Tips

1. For any questions that came up in the process of you developing your solution, please write them to a Questions.md file. Additionally, please write out your assumptions as to the answer to the question, so we can capture your logic as you proceeded from having said question.
2. Your success will be measured against your ability to:
   1. Create a durable model definition that will ultimately lead to strong data storage structures that can answer the questions presented in the Objective.
   2. Create a durable model definition that accurately represents the customer interaction and provides a pathway to answer further questions *beyond* those listed in the Objective.
   3. Explain multiple points of view for possible strategies, positioning pros against cons for a final decision.
   4. Write well suited documentation around your ideas so that it can be clearly understood when we review before the actual interview.

Please send an email to [david.sink@alaskaair.com](mailto:david.sink@alaskaair.com); [Joshua.Benard@alaskaair.com](mailto:Joshua.Benard@alaskaair.com); [karunya.namuduri@alaskaair.com](mailto:karunya.namuduri@alaskaair.com); [daidre.gothard@alskaair.com](mailto:daidre.gothard@alskaair.com) with a link to your repository when you have completed the exercise.