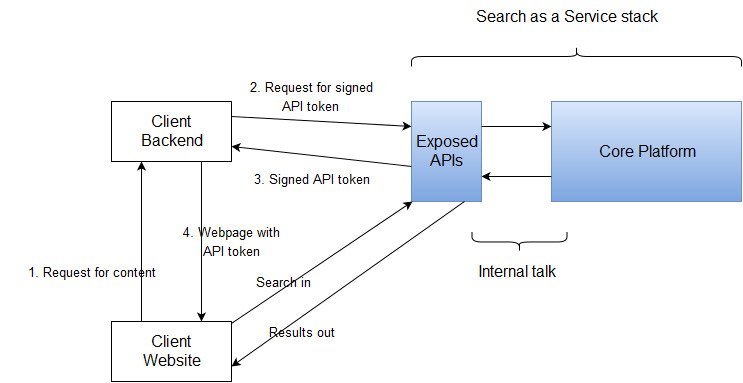
**High Level Design: Real-time Search-as-a-Service**

# Overview

Real-time Search-as-a-Service (RTS) works in the cloud using Amazon Web Services’ scalable infrastructure.

This document tries to explain the High Level Design (HLD) for the service, assuming the client to be a web app. It may be noted that the design is same for other clients, like Android app, as well.

The information flow looks roughly like this:



The basic flow can be summed up like this:

1. Client’s customer loads our client’s website.
2. In the first load, client backend makes a call to one of our APIs, asking for an API token. This request contains information like User Id, Access Control Lists (ACLs) for the user, permissions, etc.
3. Based on the parameters in the call, RTS uses HMAC SHA-256 to generate a signed API token.
4. Client’s backend embeds this token into the webpage which it shows to its customers.
5. Customers’ browser then can directly call the RTS stack for search results. RTS stack decrypts the token, and takes out all information from it, and returns the search results.
6. Token has expiry – of around 15 minutes, after which it is refreshed, and the client must call the RTS stack with the new token.