**Vehicle search service - SLO objectives**

|  |  |
| --- | --- |
| **Status** | **Published** |
| **Author** | **Sarankumar Chandrasekaran** |
| **SLO Adoption Leader** | **Balachandar Natarajan** |
| **SLI/SLO Owner** | **Saiprasad Potluri/Ratna Puttaswamy** |
| **SLI/SLO Approvers:** | **Muthupalaniappan V/Mathivanan V** |
| **Approval Date** | **10/09/2023** |
| **Revisit Date** | **01/09/2023** |
| **Ford Github URL** | **To be included <saran>** |
| **API Catalog URL** | [API Catalog Vehicle Search URL (ford.com)](https://api-viewer.apiplatform.ford.com/viewer?apiId=9532) |

# **Service Overview**

**Vehicle Search service** is designed to help users find the matching vehicles based on the input filter and sort criteria. V2 has been introduced to support budget-based search along with the existing filter based search functionality

**Vehicle Finder Inventory** provides the inventory results based on user selected vehicle configuration, payment method, dealer(s) and geo-location. This API is currently used in the model-e user journey for the NA market.

**Vehicle Search for Dealer** API provides the dealer details based on product catalog and customer selected location (radius based). In addition, also supports filter by isOrderEnrolled/isReservationEnrolled for IMG market.

**Vehicle Finder Facet** API provides the vehicle configuration facets based on given catalog, model (powertrain & series) & geo-location. This API is currently used in the model-e user journey for the NA market.

**Data** **Source**: All SLI’s mentioned in this document are observed and measured from the individual microservice’s API access log (from the Google Cloud Run layer)

**Compliance** **window**: All the service level objectives use a four-week rolling window.

# **SLI/SLO: Availability**

**Objective**: If users receive errors, incorrect data, or no data at all, the service is functionally unavailable to them. The SLO for availability aims to ensure that the service is functioning correctly and delivering the intended results to users.

**SLI**: The proportion of successful requests, as measured from the application access log metrics. Any HTTP status other than 500–599 is considered successful.

**SLI Calculation:** *count of "api" http\_requests which do not have a 5XX status code divided by count of all "api" http\_requests, on a rolling 28-day window.*

**SLO**: Percentage of successful requests. For example, an SLO of 99.5% means that at least 995 out of 1000 requests should have a non-500 response code for a rolling window of 28 days.

|  |  |
| --- | --- |
| VehicleSearch | 99.5% Success |
| VehicleSearchForDealer | 99.5% Success |
| VehicleFinderInventory | 99.5% Success |
| VehicleFinderFacets | 99.5% Success |

# **SLI/SLO: Latency**

**Objective**: The objective is to develop a system that is not only highly available but also performs effectively. Even if a system is available most of the time, if it is sluggish, users will still perceive it as unreliable. When users are required to wait for an unreasonable amount of time to receive service, it is referred to as latency. Therefore, it is important to prioritize both availability and performance to ensure that the system meets user expectations.

**SLI:** The proportion of sufficiently fast requests, as measured from the application access log metrics. “Sufficiently fast” indicates the percentage of requests that meet the performance criteria defined by the 90th and 99th percentiles of all requests served in the given measurement window.

**SLI Calculation:**

**90th Percentile measurement:** *count of number of "api" HTTP requests that have a duration less than or equal to a specified time of "XX" milliseconds and divide that by the count of all "api" HTTP requests >= 90%*

**99th Percentile measurement:** *count of number of "api" HTTP requests that have a duration less than or equal to a specified time of "XX" milliseconds and divide that by the count of all "api" HTTP requests >= 99%.*

**SLO**:

**6 hours, Multi window SLO:** This tiered SLO is based on a 6-hour multiple-window measurement. This means that the 90thand 99th percentile latency metric is less than XX ms for at least 99% of the total 6-hour windows in a 28-day compliance period. Out of the total 112 available windows of a 28-day period there can be a maximum of 1 window that may not comply with the latency requirements.

**28-day, Single window SLO**: This tiered SLO is defined for a 28-day single-window measurement. This means that all the requests received in the entire 28-day compliance window are treated as one bucket. There are no multi-window-based measurements.

|  |  |  |
| --- | --- | --- |
| VehicleSearch | 90 % of requests < 125 ms  99% of requests < 200 ms | 6 hours multi-window |
| VehicleSearchForDealer | 90 % of requests < 50 ms  99% of requests < 75 ms | Single window |
| VehicleFinderInventory | 90 % of requests < 75 ms  99% of requests < 125 ms | Single window |
| VehicleFinderFacets | 90 % of requests < 125 ms  99% of requests < 500 ms | Single window |

# **SLI/SLO: Throughput**

**Objective**: Throughput (Traffic) measures how our systems are being used. This doesn’t necessarily impact our service’s reliability. Instead, traffic is potentially a cause of other issues that impact reliability, like saturation and latency.

**SLI**: The throughput SLI specifies the number of requests that the service must be able to handle per unit of time, such as requests per second (TPS) or requests per hour (TPH).

**SLI Calculation:** *count all "api" HTTP requests that occur within a 1-second time window, regardless of their HTTP status code.*

**SLO**: The objective of the throughput Service Level Objective (SLO) is to ensure that the system is capable of handling a certain volume of requests within a specified period. The idea is to ensure that the system can handle the expected workload without experiencing performance degradation or failure.

|  |  |
| --- | --- |
| VehicleSearch | 30 TPS |
| VehicleSearchForDealer | 20 TPS |
| VehicleFinderInventory | 20 TPS |
| VehicleFinderFacets | 20 TPS |

# **Rationale**

Availability and latency SLIs were measured over a period of 28 days. The SLI/SLO mentioned in this artifact were defined by the product team and the services were verified to be running at or above those levels. The 28-day compliance period is used to ensure that the SLIs are representative of the service's overall performance. The services are verified to be running at or above the defined SLIs by using a variety of monitoring tools and techniques.

# **Error Budget**

Each objective has a separate error budget, which is defined as the difference between 100% and the goal for that objective. For example, if the goal for API availability is 99.5%, then the API availability error budget is 0.5%.

The error budget is used to determine how many errors are allowed before the service is considered to be in violation of its SLO. For example, if the API availability error budget is 5 errors per 1000 requests, then the service is considered to be in violation of its SLO if there are more than 5 errors in a given time period.

The error budget will be used to help teams prioritize their work and make decisions about how to allocate resources. For example, if a team has a limited budget for fixing errors, they can use the error budget to determine which errors are the most important to fix.

# **Clarifications and Caveats**

* The latency metrics for the vehicle search service were defined by the product team based on the criticality and dependencies of the API. The product team determined which latency metrics were most important to measure and how they would be used to track the performance of the API.
* The product team also defined the error codes that would be counted as errors. HTTP 5XX status messages are considered to be errors because they indicate that the API was not able to process the request successfully. All other status codes are considered to be successes.
* The availability and latency metrics for the vehicle search service were determined after considering the SLA for the various Google managed cloud services. Where necessary, the SLO metrics for the appropriate upstream dependent systems were also taken into account before arriving at the target SLO for vehicle search service.

**SLO Exclusions**

The SLO does not apply to certain features and circumstances, including:

1. Features that have been designated as pre-general availability (Alpha, Beta), as they are still in development and may not yet meet the standards of the SLO.
2. Errors that are caused by our Ford internal or external 3rd party integrations, as well as upstream systems. These errors are beyond our direct control and may have a ripple effect on the performance of the system, making it difficult to guarantee the terms of the SLO.
3. External factors that are outside of our control, such as issues with Google’s compute layer.

These factors will impact the performance of the system and fall outside of the scope of the SLO.