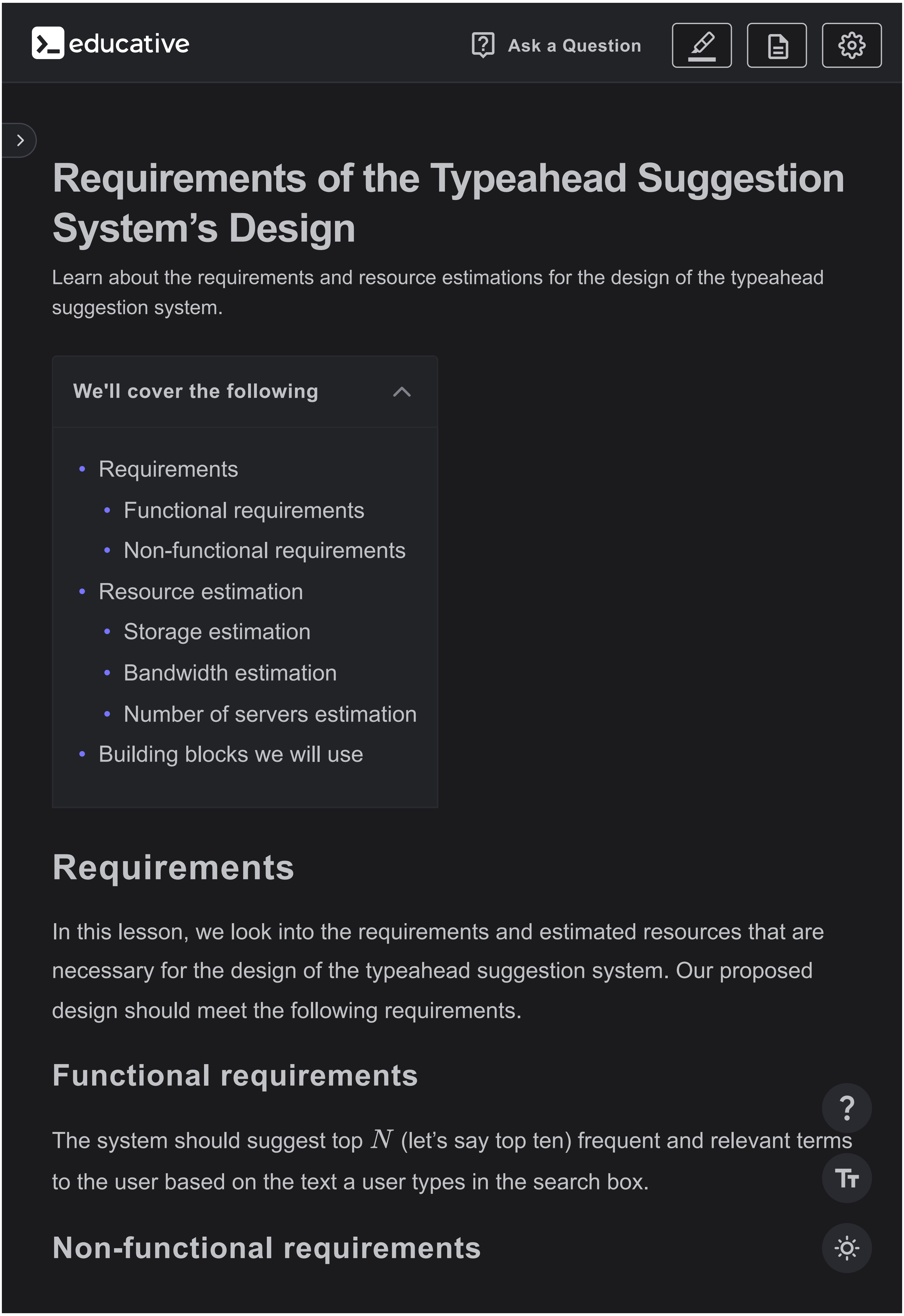


Requirements of the Typeahead Suggestion System’s Design

Learn about the requirements and resource estimations for the design of the typeahead suggestion system.

We'll cover the following

Requirements

In this lesson, we look into the requirements and estimated resources that are necessary for the design of the typeahead suggestion system. Our proposed design should meet the following requirements.

Functional requirements

The system should suggest top N (let’s say top ten) frequent and relevant terms to the user based on the text a user types in the search box.

Non-functional requirements

- **Low latency:** The system should show all the suggested queries in real time after a user types. The latency shouldn’t exceed 200 ms. A study suggests that the average time between two keystrokes is 160 milliseconds. So, our time-budget of suggestions should be greater than 160 ms to give a real-time response. This is because if a user is typing fast, they already know what to search and might not need suggestions. At the same time, our system response should be greater than 160 ms. However, it should not be too high because in that case, a suggestion might be stale and less useful.
- **Fault tolerance:** The system should be reliable enough to provide suggestions despite the failure of one or more of its components.
- **Scalability:** The system should support the ever-increasing number of users over time.

Resource estimation

As was stated earlier, the typeahead feature is used to enhance the user experience while typing a query. We need to design a system that works on a scale that’s similar to Google Search. Google receives more than 3.5 billion searches every day. Designing such an enormous system is a challenging task that requires different resources. Let’s estimate the storage and bandwidth requirements for the proposed system.

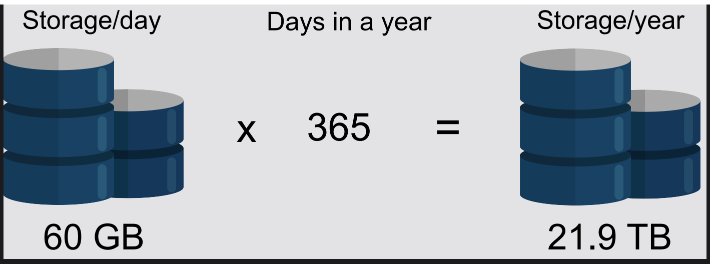
Storage estimation

Assuming that out of the 3.5 billion queries per day, two billion queries are unique and need to be stored. Let’s also assume that each query consists of 15 characters on average, and each character takes 2 Bytes of storage. According to this formulation, we would require the following:

$2\ billion\ \times 15 \times 2 = 60GB$ to store all the queries made in a day.

Storage required per year:

$$60GB/day \times 365 = 21.9TB/year$$



Storage requirements for the typeahead suggestion system

Bandwidth estimation

3.5 billion queries will reach our system every day. Assume that each query a user types is 15 characters long on average.

Keeping this in mind, the total number of reading requests of characters per day would be as follows:

$15 \times 3.5\ billion = 52.5\ billion$ characters per day.

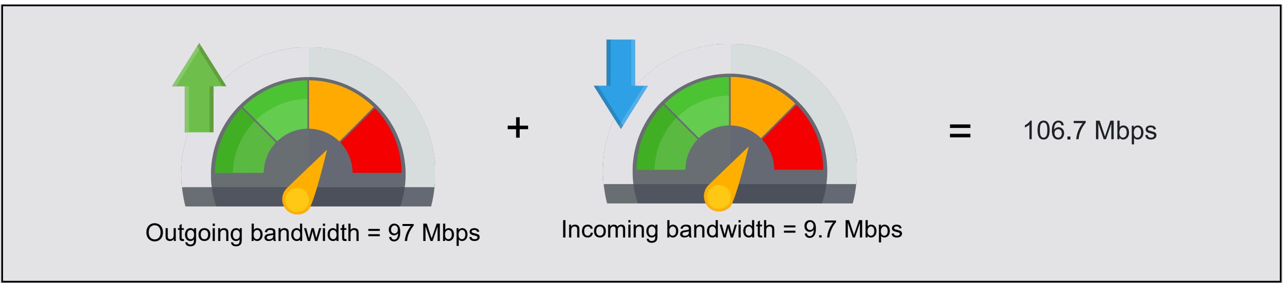
Total read requests per second: $52.5B/86400 \approx 0.607M$ characters/sec. 86,400 is the number of seconds per day.

Since each character takes 2 Bytes, the bandwidth our system would need is as follows:

$0.607M \times 2 \times 8 = 9.7Mb/sec$

$9.7Mb/sec$ is the incoming bandwidth requirement for queries that have a maximum length of 15 characters. Our system would suggest the top ten queries that are roughly of the same length as the query length after each character a user types. Therefore, the outgoing bandwidth requirement would become the following:

$10 \times 9.7Mb/sec = 97Mb/sec.$



The total bandwidth required by the typeahead suggestion system

Number of servers estimation

Our system will receive 52.5 billion requests per day. Considering our assumption of using daily active users as a proxy for the number of requests per second for peak load, we get 52.5 billion requests per second (considering each character as a request).

Recall that a typical server can serve 64,000 requests per second (RPS). So, the required servers can be estimated using the following formula:

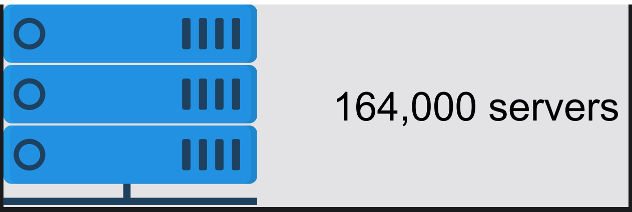
$$Number\ of\ Servers\ = \frac{Number\ of\ requests/second}{RPS\ of\ server}$$

$$Number\ of\ Servers\ = \frac{52.5\ billion}{64,000} = 820312.5 \approx 821K\ servers$$

> The initial estimate of 52.5 billion requests per second seems unrealistic. However, on average, a person types at a rate of around three to four characters per second.

Therefore, assuming 3.5 billion users and a peak load scenario where each user types roughly three characters per second, we arrive at a more feasible estimate of 10.5 (3.5×3) billion requests per second. This revised figure provides a more realistic estimation as follows:

$$Number\ of\ Servers\ = \frac{3.5\ billion \times 3}{64,000} = 164062.5 \approx 164K\ server$$



The number of servers required for the typeahead suggestion system

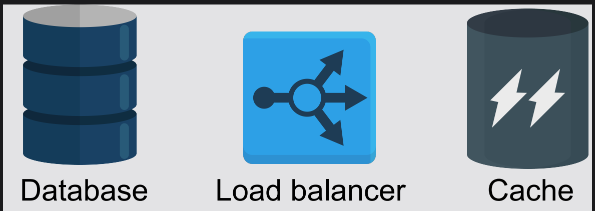
In the table below, adjust the values to see how the resource estimations change.

Resource Estimation for the Typeahead Suggestion System

Total Queries per Day	3.5	billion
Unique Queries per Day	2	billion
Minimum Characters in a Query	15	Characters
Server's QPS	64000	Queries per second
Storage	*f* 60	GB/day
Incoming Bandwidth	*f* 9.7	Mb/sec
Outgoing Bandwidth	*f* 97	Mb/sec
Number of Servers (uniform requests)	*f* 9	Severs
Number of Servers (peak load)	*f* 164063	Servers

Building blocks we will use

The design of the typeahead suggestion system consists of the following building blocks that have been discussed in the initial chapters of the course:



Building blocks required in the design of the typeahead suggestion system

- **Databases** are required to keep the data related to the queries’ prefixes.
- **Load balancers** are required to disseminate incoming queries among a number of active servers.
- **Caches** are used to keep the top N suggestions for fast retrieval.

In the next lesson, we’ll focus on the high-level design and APIs of the typeahead suggestion system.

← Back

☑ Completed

Next →