



Data Ingestion Use Cases

In this lesson, we will discuss some common data ingestion use cases in the industry.

We'll cover the following



- Moving Big Data into Hadoop
- Streaming data from databases to Elasticsearch server
- Log processing
- Stream processing engines for real-time events

This is the part where I talk about some of the data streaming use cases commonly required in the industry.

Moving Big Data into Hadoop#

This is data ingestion's most popular use. As discussed before, Big Data from IoT devices, social apps, and other sources, streams through data pipelines and moves into the most popular distributed data processing framework Hadoop for analysis.



Streaming data from databases to Elasticsearch server#

Elastic search is an open-source framework for implementing search in web applications. It is a de facto search framework used in the industry simply because of its advanced features and it being open source. These features enable businesses to write their own custom solutions when they need them.

In the past, I wrote a product search software with a few of my friends as a service using *Java*, *Spring Boot*, and *Elastic search*. We would stream and index quite a large amount of product data from the legacy storage solutions to the Elastic search server in order to make the products come up in the search results.

All the data intended to show up in the search was replicated from the main storage to the Elastic-search storage. Also, as the new data was persisted in the main storage it was asynchronously delivered to the Elastic server in real-time for indexing.

Log processing#

If your project isn't a hobby project, chances are it's running on a cluster. When we talk about running a large-scale service, monolithic systems are a thing of the past. With so many microservices running concurrently. There is a massive number of logs, which are generated over a period of time. Logs are the only way to move back in time, track errors, and study

the system's behavior.



So, to study the behavior of the system holistically, we have to stream all the logs to a central place. Ingest logs to a central server to run analytics on it with the help of solutions like the Elastic LogStash Kibana(ELK) stack, etc.

Stream processing engines for real-time events#

Real-time streaming and data processing are the core components in systems handling LIVE information such as sports. It's imperative that the architectural setup in place is efficient enough to ingest data, analyze it, figure out the behavior in real-time, and quickly push the updated information to the fans. After all, the whole business depends on it.

Message queues like *Kafka* and stream computation frameworks like *Apache Storm*, *Apache Nifi*, *Apache Spark*, *Samza*, *Kinesis*, etc are used to implement the real-time large-scale data processing features in online applications.

This is a good read on the topic:

An Insight into Netflix's real-time streaming platform
(<https://medium.com/netflix-techblog/keystone-real-time-stream-processing-platform-a3ee651812a>)


Alright!! time to have a look into data pipelines in the next lesson.

Different Ways of Ingesting Data and ...



Data Pipelines

 Mark as Completed

 Report an Issue