Explain Logstash, its pipeline stages, and common input, filter, and output plugins

🔹 **Short Interview Answer**

Logstash is a data pipeline tool in the ELK stack. It has three stages:

- **Input** (collects data from sources like Beats, syslog, files),
- **Filter** (transforms data using plugins like grok, mutate, date), and
- **Output** (sends data to Elasticsearch, stdout, or other systems).

🔹 **What is Logstash?**

- **Logstash** is a **data processing pipeline tool** in the ELK stack.
- It **collects data** (logs, metrics, events) → **processes/transforms** it → **sends it to a destination** (like Elasticsearch).
- It's flexible: supports 200+ plugins.

👉 Think of Logstash as a **water treatment plant** 💧:

- **Input** = raw water coming in.
- **Filter** = cleaning, purifying, transforming water.
- **Output** = clean water delivered to homes.

🔹 **Example Raw Log (messy)**

2025-09-17 10:22:33,456 ERROR User login failed for user=JohnDoe ip=192.168.1.10

This is just a text line. Hard for Elasticsearch to search effectively.

🔹 **Cleaning**

- Remove unnecessary spaces/symbols
- Standardize the timestamp format
- Drop duplicate or empty fields

After cleaning:

timestamp="2025-09-17T10:22:33Z" level="ERROR" message="User login failed" user="JohnDoe" ip="192.168.1.10"

🔹 **Transforming**

- Extract fields into **key-value pairs**
- Change data types (e.g., ip as IP field, timestamp as datetime)

- Mask sensitive info (e.g., replace username with anon_user)
- Add new info (e.g., geo-location from IP, hostname, or tags)

After transformation:

```
{
 "timestamp": "2025-09-17T10:22:33Z",
 "level": "ERROR",
 "event": "login_failed",
 "user": "anon_user",
 "ip": "192.168.1.10",
 "geo_location": "Bangalore, India"
}
```

🔹 Why is this useful?

- Now in Elasticsearch, you can **search and filter** easily:
  - Show all logs where event=login_failed
  - Find logs from geo_location = Bangalore
  - Count how many ERROR events happened in the last 1 hour

⚡ In short:

- **Clean = make logs neat and consistent**
- **Transform = enrich or restructure logs into useful fields**

🔹 Logstash Pipeline Stages

1. **Input** 🛠️
   - Where data comes from.
   - Example: log files, syslog, Beats agents.
2. **Filter** 🔄
   - Transform/manipulate data.
   - Example: parse timestamps, extract fields, drop unwanted info.
3. **Output** 📤
   - Where processed data is sent.
   - Example: Elasticsearch, Kafka, stdout (console).

🔹 Common Logstash Input Plugins

- **Beats** → Collects logs/metrics from servers (Filebeat, Metricbeat).
- **Syslog** → Ingests system logs from Linux/Unix servers.
- **File** → Reads local log files (e.g., /var/log/nginx/access.log).
- **Kafka** → Streams logs from Kafka topics.

👉 Example:

```
input {
 beats { port => 5044 }
}
```

🔹 Common Logstash Filter Plugins

- **Grok** → Pattern matching, parses unstructured logs into structured fields.

    - Example: Extract IP, status code, URL from Apache log.
- **Mutate** → Rename, remove, modify fields.
- **Date** → Parse date strings into standard timestamp.
- **GeoIP** → Add geolocation info from an IP address.

👉 Example:

```
filter {
  grok { match => { "message" => "%{IP:client} %{WORD:method} %{URIPATH:request}" } }
  date { match => ["timestamp", "ISO8601"] }
}
```

🔷 **Common Logstash Output Plugins**

- **Elasticsearch** → Store data for searching & visualization.
- **Stdout** → Print to console (for debugging).
- **File** → Write to local file.
- **Kafka** → Send processed events to Kafka for further use.

👉 Example:

```
output {
  elasticsearch { hosts => ["http://localhost:9200"] }
  stdout { codec => rubydebug }
}
```

Basics

What is **Logstash** in the ELK stack?

A) A visualization tool
B) A search engine
C) A data processing pipeline
D) A monitoring dashboard

Answer: C) A data processing pipeline

Pipeline Stages

Which of the following correctly represents the **Logstash pipeline stages**?

A) Input → Transform → Query
B) Collect → Analyze → Store
C) Input → Filter → Output
D) Source → Index → Visualize

Answer: C) Input → Filter → Output

If you want Logstash to read logs directly from /var/log/nginx/access.log, which input plugin would you use?

A) Beats
B) File
C) Syslog
D) Kafka

Answer: B) File

Which Logstash filter plugin is commonly used to **parse unstructured logs** (like Apache logs) into structured fields?

A) Mutate
B) Date
C) GeoIP
D) Grok

Answer: D) Grok

You want to **rename and remove fields** in incoming logs. Which filter plugin is best suited?

A) Grok
B) Mutate
C) GeoIP
D) Date

Answer: B) Mutate

If you want to **add location details based on an IP address**, which filter plugin should you use?

A) Date
B) GeoIP
C) Grok
D) Mutate

Answer: B) GeoIP

Which Logstash **output plugin** sends data to Elasticsearch for storage and visualization?

A) File
B) Stdout
C) Kafka
D) Elasticsearch

Answer: D) Elasticsearch
