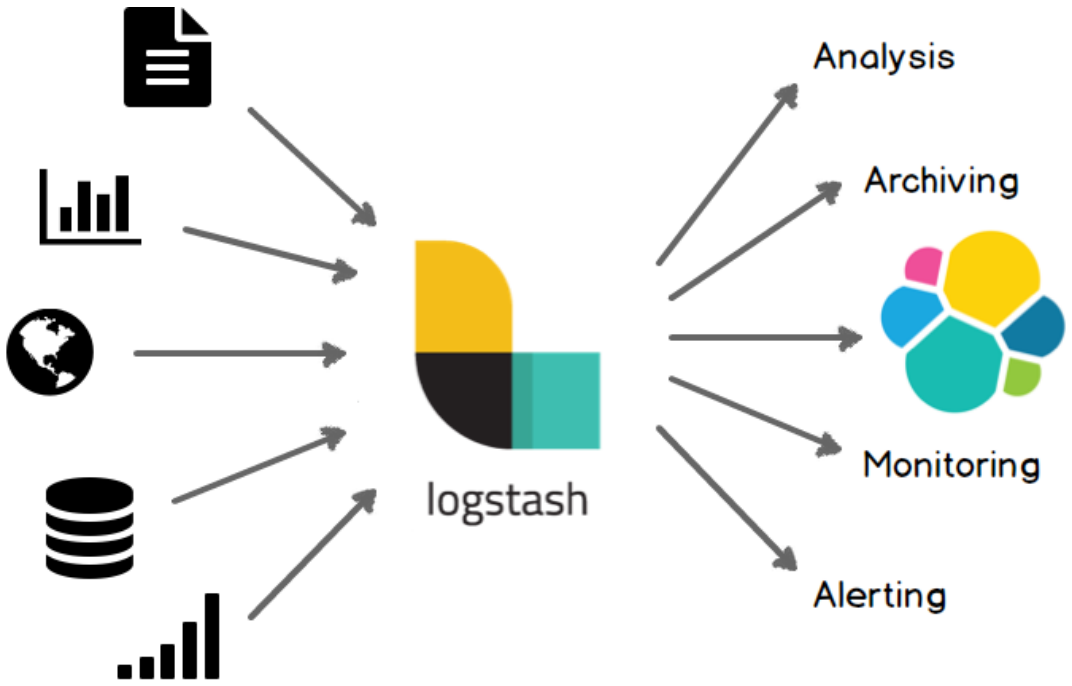
Elasticsearch是当前主流的分布式大数据存储和搜索引擎，可以为用户提供强大的全文本检索能力，广泛应用于日志检索，全站搜索等领域。

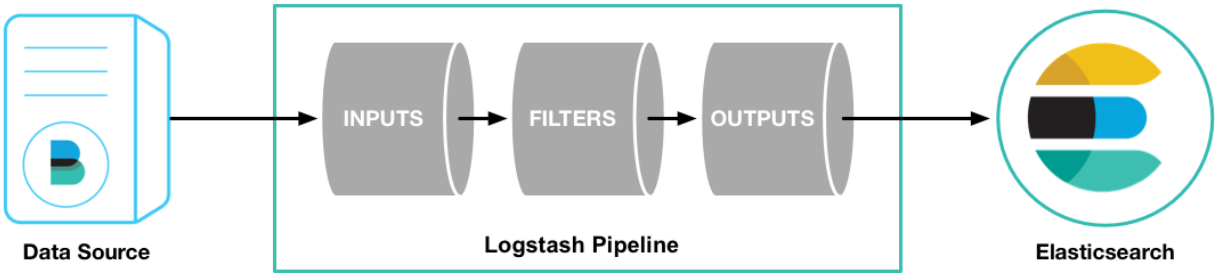
Logstash作为Elasicsearch常用的实时数据采集引擎，可以采集来自不同数据源的数据，并对数据进行处理后输出到多种输出源，是Elastic Stack 的重要组成部分。

本文从Logstash的工作原理，使用示例，部署方式及性能调优等方面入手，为大家提供一个快速入门Logstash的方式。文章最后也给出了一些深入了解Logstash的的链接，以方便大家根据需要详细了解



1 Logstash工作原理

1.1 处理过程



如上图，Logstash的数据处理过程主要包括：**Inputs, Filters, Outputs** 三部分， 另外在Inputs和Outputs中可以使用**Codecs**对数据格式进行处理。这四个部分均以插件形式存

在，用户通过定义pipeline配置文件，设置需要使用的input，filter，output, codec插件，以实现特定的数据采集，数据处理，数据输出等功能

- （1）Inputs：用于从数据源获取数据，常见的插件如file, syslog, redis, beats等
- （2）Filters：用于处理数据如格式转换，数据派生等，常见的插件如grok, mutate, drop, clone, geoip等
- （3）Outputs：用于数据输出，常见的插件如elastcisearch，file, graphite, statsd等
- （4）Codecs：Codecs不是一个单独的流程，而是在输入和输出等插件中用于数据转换的模块，用于对数据进行编码处理，常见的插件如json，multiline

可以点击每个模块后面的详细参考链接了解该模块的插件列表及对应功能

1.2 执行模型：

- （1）每个Input启动一个线程，从对应数据源获取数据
- （2）Input会将数据写入一个队列：默认为内存中的有界队列（意外停止会导致数据丢失）。为了防止数丢失Logstash提供了两个特性：
  Persistent Queues：通过磁盘上的queue来防止数据丢失
  Dead Letter Queues：保存无法处理的event（仅支持Elasticsearch作为输出源）
- （3）Logstash会有多个pipeline worker, 每一个pipeline worker会从队列中取一批数据，然后执行filter和output（worker数目及每次处理的数据量均由配置确定）