- Datasource : source of the data (taxi data)
- data aggregators enables data aggregation services to collect data through any source channels and make it useable.

Fig1:

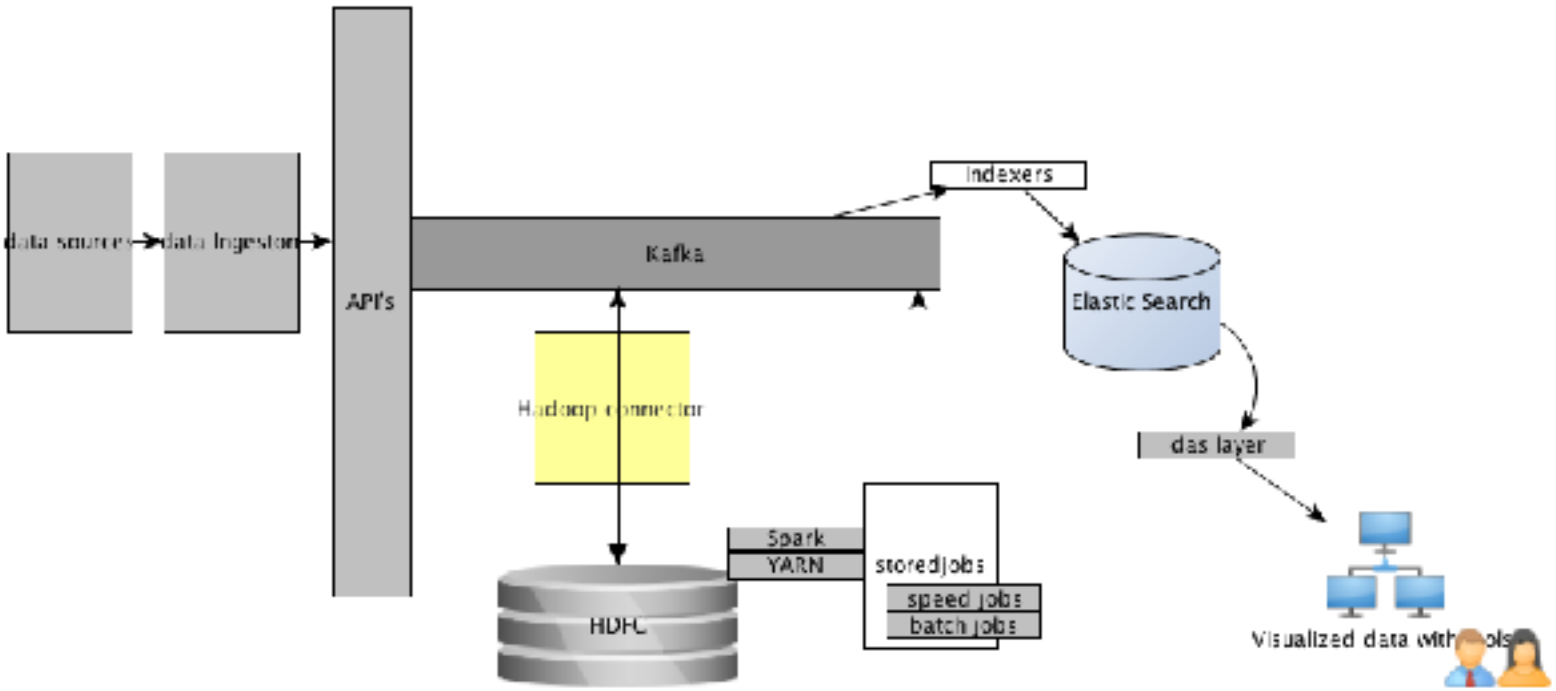
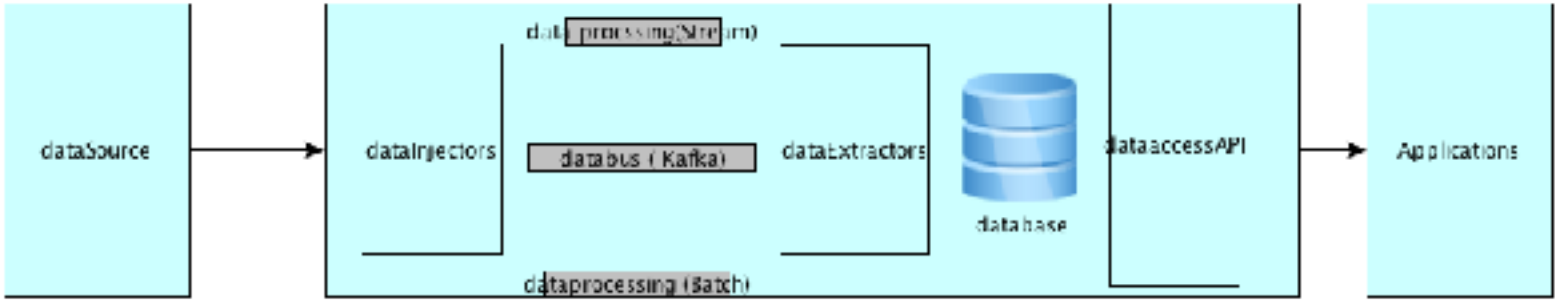


Fig2 :



- Hadoop connector gets files from or sends files to data directories on the Hadoop Distributed File System (HDFS) .
- Batch jobs are executing on historical data based on Predefined intervals.
- Speed job is executed on the streaming data.

- Kafka is a distributed publish-subscribe messaging system - designed to be fast, scalable, and durable. Kafka maintains messages in topics. Producers write data to topics and consumers read from topics.

- Queuing systems, such as Kafka, provides the ability to keep the messages. If streaming process has some delay, or the Spark cluster goes down, we will not have any data message loss because they are all stored in a queuing system.

- (processed data will push to Elastic Search.)database and a serving layer where it can be another API or a frontend web app. or Visualize with any Visualization tools tableau etc…