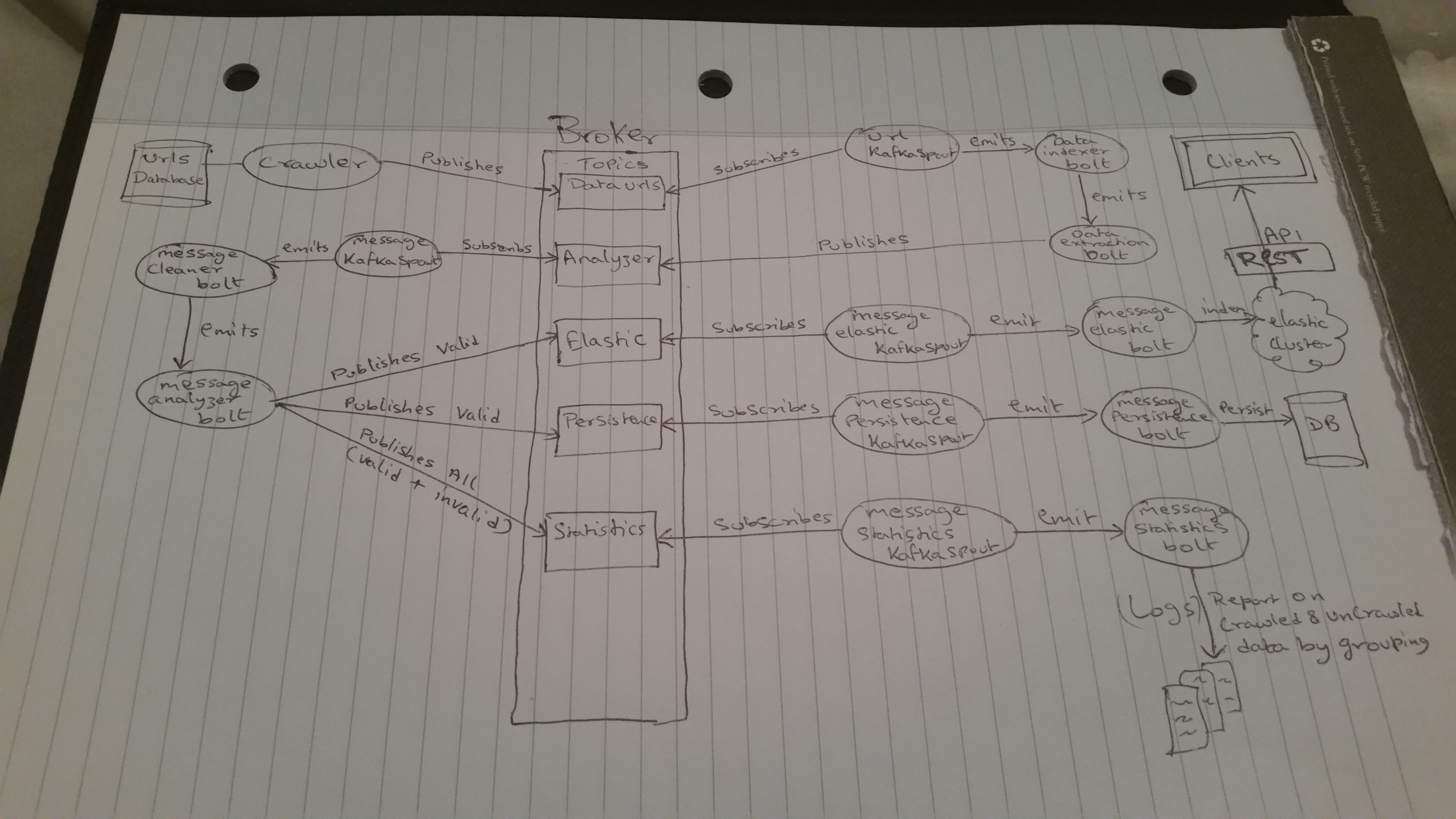
**Overview:**

**NOTES:** This architecture is **Single broker – Multiple topics – One partition/topic;** can be extended to multiple brokers and consumer groups by adding in few configurations.



**Services:**

**Crawler Service**

**Function:** Crawls website to scrap raw product data (Crawler referenced by feed name”retailer name”)

**Detail:** Contains one spout and two bolts – URL KafkaSpout, Data Indexer bolt and Data Extraction Bolt

*URL Spout* –Subscribes to Kafka topic (Data URLs), consumes product URL’s extracted by the crawler and emits them to Data Indexer bolt

*Data Indexer Bolt* – Consumes URL’s emitted by URL Spout and indexes them in database, emits unique product URL’s to Extract bolt.

*Data Extraction Bolt* – Scraps the website to collect raw product data and publishes raw data in to Kafka topic ”Analyzer”.

**Analyzer Service**

**Function:** Cleans and analyzes raw product data for its validity to satisfy the rules of quality data as specified.

**Details:** Contains one spout and two bolt – Message KafkaSpout, Message Cleaner Bolt and Message Analyzer Bolt

*Message KafkaSpout* – Subscribes to kafka topic (Analyzer), consumes product data from and emits them to Message Cleaner Bolt.

*Message Cleaner Bolt* – Cleans the data (Like: URL’s, Price, Remove Special Characters from product description etc.) and emits the cleaned data to Message Analyzer Bolt.

*Message Analyzer Bolt* –Analyzes for its quality and publishes the only valid product data to 2 Kafka topic’s (Elastic, Persistence), publishes valid & invalid data to Kafka topic (Statistics)

**{**

*Elastic* Topic is a queue for indexing data in to elastic cluster

*Persistence* Topic is a queue for indexing data in to persistent database.

*Statistics* Topic is a queue for logging information on products crawled successfully and failed.

**}**

**Persistence Service**

**Function**: Stores the crawled products in database

**Details**: Contains one spout and one bolt – Message Persistence KafkaSpout and Message Persistence Bolt

*Message Persistence KafkaSpout* –Subscribes to Kafka topic (Persistence), consumes product data from topic and emits them to Message Persistence Bolt

*Message Persistence Bolt* – Consumes product data emitted by Persistence KafkaSpout and persist them into database.

**Elastic Search Service**

**Function**: Stores crawled products in elastic search cluster

**Details**: Contains one spout and one bolt – Message Elastic KafkaSpout and Message Elastic Bolt

*Message Elastic KafkaSpout* - Subscribes to Kafka topic (Elastic), consumes product data from topic and emits them to Message Elastic Bolt

*Message Elastic Bolt* – Consumes product data emitted by Elastic KafkaSpout and indexes them into elastic search cluster.

**Logging Service**

**Function**: Gives statistics on crawled data – Failures and Success for each retailer.

**Details**: Contains one spout and one bolt – Message Logger KafkaSpout and Message Logger Bolt

*Message Logger KafkaSpout-* Subscribes to Kafka topic (Statistics), consumes product data from topic and emits them to Message Logger Bolt

*Message Logger Bolt* – Consumes product data emitted by Logger KafkaSpout and groups them by retailers.

**Product REST API**

**Function**: Exposes Restful API to access elastic cluster.

**Details**: Exposes search endpoints to access product information from elastic cluster (on the backend elastic search performs full text search (precision, proximity search, partial matching using analysers– based on search request), filter, paginates and returns the actual or closest results)