# 概要设计

## 技术概念

### Elasticsearch

Elasticsearch提供了一个支持高并发、高可用、易扩展的分布式搜索存储解决方案，restful风格的设计，能够提供易用的查询和共享接口。它在解决大数据相关问题中的良好表现，使得它成为大数据解决方案中的首选工具之一，下面是它的主要特性：

1. 多租户支持。在共用的数据中心内，以单一架构和服务提供多个用户使用，并提供定制化的服务。Elasticsearch支持多个索引间的数据隔离，从底层支持了多租户技术。
2. 分布式。当多个节点存在同一网段下，设置相同的集群名称，即可组成一个分布式集群。良好的横向扩展设计，在数据规模的快速增长的情况下，可以弹性地增加服务器资源。
3. 高可用性。Elasticsearch采用无中心化的集群设计方式，集群中的中央节点出现故障后，elasticsearch会自动甄选出一个新的节点作为中央节点，并自动进行数据分片的转移，保证用户数据的安全与访问。
4. 全文搜索。Elasticsearch底层采用lucene实现，支持lucene所有的功能。
5. 索引机制。Elasticsearch创建的索引包含以下的概念：索引、索引机制、Mapping、数据类型。与关系型数据库相比较，索引相当于数据库，索引类型相当于数据表，mapping相当于表结构，数据类型相当于列。单个索引中可以存在多个索引类型，索引类型由基本数据类型构成。Elasticsearch采用的索引机制为异构数据的存储提供了新思路。
6. Restful风格的API。REST是以资源为中心，把web看作一系列具有不同具体表现形式的抽象资源的组成，利用uri来标识资源，并提供HTTP协议形式的统一访问接口。利用restful架构，elasticsearch可以实现资源的封装，并利用rest接口提供共享和检索服务。

### Elasticsearch-sql

Elasticsearch-sql提供了对sql语句的支持，提供一个web页面，可以很方便的执行sql查询，主要面向管理人员，提高常见操作的使用效率。

## 系统模型设计

如图一所示，基于elasticsearch的数据存储与搜索系统主要包括元数据存储模型、元数据信息抽取模型、创建索引模型、查询转换模型以及搜索统计服务模型等。

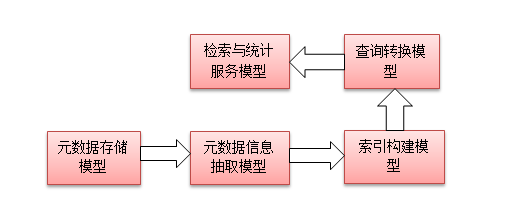


图1 系统模型

元数据存储模型为数据存储与搜索平台的数据源，提供文本导入、数据库导入等获取元数据的方法，文本导入为系统提供文本文件数据的导入接口，数据库导入为系统提供数据库数据同步的接口；元数据信息抽取模型用于从平台存储的元数据抽取可以标识元数据的核心信息；索引构建模型对导入的元数据信息构建分布式索引；查询转换模型可以将sql语句转换成elasticsearch的DSL查询语句，减少用户的学习成本；检索与统计服务模型提供了检索与数据统计的功能。

### 元数据信息抽取

在各类业务系统中，存在大量的异构元数据，主要表现为元数据数据格式的不统一，针对异构数据，需要对数据进行按类别进行数据抽取。系统在文本数据导入时提供数据自动识别，根据规则库对数据进行自动匹配识别，并提供直接正则方法进行字段抽取并将数据导入系统；在数据数据同步时，提供相应的配置，抽取数据库字段，并将数据同步到elasticsearch。

### 构建索引

由于 ElasticSearch 支持分布式索引服务， 对于使用者来说，只需要调整索引存储节点的分配策略即可实现系统的分布式运行要求，因而构建索引的问题集中到了异构元数据的处理上，默认情况下，索引类型中的数据类型主要由数据源的数据类型来确定，这也会导致索引数据类型可能与元数据类型不一致的情况，所以，这里提供默认配置导入以及手工设置类型的mapping来保证索引类型与元数据的类型能够一一对应。ElasticSearch 创建的索引在逻辑上是隔离的，不同用户创建的索引文件互不干扰， 为用户之间的数据共享提供了支撑。ElasticSearch 是面向JSON文档进行操作的，平台通过把抽取的元数据信息转换为JSON格式的文档， 利用创建好的Mapping对JSON文档进行匹配来创建索引。

### 查询转换

数据导入平台成功后，查询是一个最直接的需求，但是es自身的查询语句DSL却有着一定的学习成本，不能快速上手进行查询数据，因此系统提供了sql语句查询入口，方便用户对数据进行查询，提高使用效率。

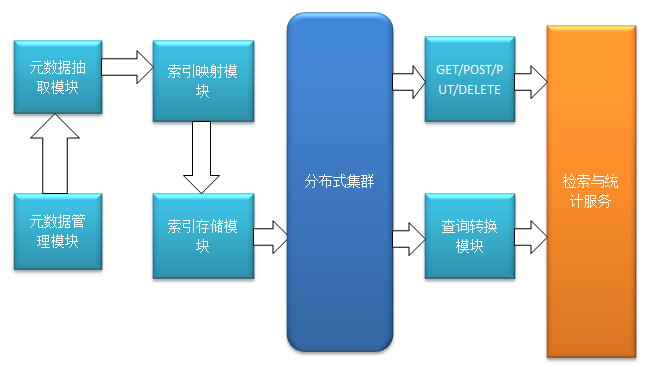
### 检索与统计模型

ElasticSearch 采用了Restful设计风格，其创建的所有索引资源使用了统一资源标识符(URI) 进行统一描述。ElasitcSearch 提供了GET、POST、PUT、DELETE 四种动作来操作索引资源， 其中GET方法用于查询索引资源，POST方法用于更新索引资源，PUT方法用于向索引文件中增加记录，DELETE 方法用于删除索引资源。

此外，elasticsearch还提供了聚合来实现统计分析的功能，通过聚合中bucket（桶，满足某个条件的文档集合）与Metrics（指标，为某个bucket中的文档计算得到的统计信息）实现统计分析的功能。sql语句示例SELECT COUNT(color) FROM table GROUP BY color，COUNT(color)就相当于Metrics，GROUP BY color则相当于一个bucket。

## 存储与搜索系统实现

依据上述系统模型实现了基于ElasticSearch的元数据搜索与共享平台，平台主要由元数据管理模块、元数据抽取模块、索引映射模块、索引存储模块、分布式集群、查询转换模块、检索与统计服务模块构成，各模块之间的关系如下图所示：



系统模块与关系图

1. 元数据管理模块负责数据源的管理
2. 元数据抽取模块负责数据字段的提取等功能
3. 索引映射模块负责索引类型的mapping设置功能
4. 索引存储模块负责索引的存储
5. 分布式集群由es集群构成，组成分布式索引集群。
6. 查询转换模块负责将sql语句转换成DSL语句并执行
7. GET/POST/PUT/DELETE提供对数据的操作
8. 检索与统计服务提供数据的搜索，以及数据的统计功能，并将结果展示给用户。