# 技术文件

## 项目总体设计

### 大数据技术架构及主体应用技术架构

1、实施方案（包含但不限于项目目标、实施方案、组织计划、技术力量投入等，具体要求详见“第五章需求及技术要求”）。

2、设备主要技术指标、参数及性能的详细说明，投标人须提供包含相关指标项的证明材料，证明材料可以使用生产厂家官方网站截图或产品白皮书或第三方机构检验报告或其他相关证明材料。未提供有效证明材料或证明材料中内容与所填报指标不一致的，该指标按不满足处理。

#### WEB页面开发技术

web页面开发技术基于vue框架，es6，nodejs，webpack，npm等技术，结合nginx，完成前端页面的开发，及访问的负载均衡。Vue是一套用于构建用户界面的渐进式[JavaScript](https://baike.baidu.com/item/JavaScript/321142" \t "https://baike.baidu.com/item/Vue.js/_blank)框架。与其它大型框架不同的是，Vue 被设计为可以自底向上逐层应用。Vue 的核心库只关注视图层，方便与第三方库或既有项目整合。Es6的全称是ECMAScript 6.0，它是JavaScript语言的下一代标准，已经在2015年6月正式发布了。它的目标，是使得JavaScript语言可以用来编写复杂的大型应用程序，成为企业级开发语言。Node.js 是一个基于 Chrome V8 引擎的 JavaScript 运行环境。 Node.js 使用了一个事件驱动、非阻塞式 I/O 的模型。Node 是一个让 JavaScript 运行在服务端的开发平台，它让 JavaScript 成为与PHP、Python、Perl、Ruby 等服务端语言平起平坐的脚本语言。webpack是一个现代JavasScript应用程序的模块打包器(module bunder)，它将sass/less 等预编译的css语言转换成浏览器识别的css文件，并搭建开发环境开启服务器，监视文件改动，热部署，将单文件组件(\*.vue)类型的文件，转化成浏览器识别的内容

#### 数据抽取转换加载技术

基于Python，java语言完成hbase数据和elasticsearch数据的抽取转换，其中使用redis作为数据调度器，hbase为数据源，它作为Hadoop生态环境的一部分，担任着文本存储的角色，弥补了hdfs只适合存储大文件的缺点；elasticsearch为数据存储仓库，完成对hbase数据的二次索引建立，最终借助elasticsearch为舆情系统提供全文检索功能。

#### 全文检索技术

基于elasticsearch搜索系统进行数据的全文检索，统计分析，相似查询等功能。ElasticSearch是一个基于[Lucene](https://baike.baidu.com/item/Lucene/6753302" \t "https://baike.baidu.com/item/elasticsearch/_blank)的搜索服务器。它提供了一个分布式多用户能力的全文搜索引擎，基于RESTful web接口。Elasticsearch是用Java语言开发的，并作为Apache许可条款下的开放源码发布，是一种流行的企业级搜索引擎。ElasticSearch用于[云计算](https://baike.baidu.com/item/%E4%BA%91%E8%AE%A1%E7%AE%97/9969353" \t "https://baike.baidu.com/item/elasticsearch/_blank)中，能够达到实时搜索，稳定，可靠，快速，安装使用方便。官方客户端在Java、.NET（C#）、PHP、Python、Apache Groovy、Ruby和许多其他语言中都是可用的。根据DB-Engines的排名显示，Elasticsearch是最受欢迎的企业搜索引擎，其次是Apache Solr，也是基于Lucene。

#### 视频处理技术

视频处理是基于视频特征提取比对和关键帧提取这两项关键技术完成的。

其中视频特征提取主要目的是进行视频对比，即视频相似度的计算。进而应用于视频检索系统。

视频的特征提取可以分为声音的特征特提取和图像的特征提取（基于ffmpeg开源库抽取关键帧）， 特征提取是一个关键的步骤，为后面机器学习算法的应用提供了基础。

视频中的帧特征提区采用基于关键帧提取技术，将视频关键帧提取出来后，基于关键帧序列，分割视频镜头。我们认为一个镜头中的视频帧具有大致相同的图片内容，同时具有大致相同的图片特征，基于这种思想，我们将一个镜头的关键帧特征融合在一起，提取一个特征向量。整个视频用各个镜头的特征来表达，同时根据不同镜头的时长作为镜头的权重。

视频图像特征对比就是根据视频镜头图像特征及镜头时长权重为依据，计算镜头间图片特征相似度并乘上相应的权重。

#### 自然语言处理技术

自然语言处理（natural language processing ，NLP）技术是人工智能领域的重要分支之一。自然语言处理技术分为自然语言理解和自然语言生成。

自然语言理解从语言的篇，章，节，段，句，词，字不同层次进行研究，为语义理解提供技术基础。自然语言处理的基础技术包括语种判别，词例化（汉语分词），词性标注，词干提取，命名实体识别，关键词提取，分句等技术。自然语言生成技术随着人工智能的发展逐步成为研究热点。

自然语言处理技术在实际应用中提供众多基础技术基础，如机器翻译系统，全文检索系统，专家系统等。

#### 文本特征提取技术

自然语言处理中，需要将文字序列的文本转变成计算机易于处理的向量数据，通过文本向向量数据的映射，用向量数据表示文本。文本向向量数据映射的过程就是文本特征提取的过程。

基础的文本特征基于词语的表示。对文本中出现的词的统计，词出现标为1，否则为0，即词袋法表示文本特征。在此基础上进行改进，提出了tf-idf特征。tf-idf特征，统计词频和到文档频率对词袋特征进行改进，充分考虑高频词和通用词。上述方法提取的文本特征随词量的增加，维度显著增加，计算耗时显著增加，通过svd分解技术实现特征维度约简，有效减少特征维度。

随深度学习技术的快速发展，深度特征被广泛应用。现有的技术有基于词的特征向量化和基于文本的特征向量化，如gensim库的word2vect和doc2vect，通过训练集训练生成模型，载入词或文本，直接映射到多为向量空间中，实现文本特征提取。

#### 文本关键词提取技术

关键词提取是文本表示的基础技术，通过展示关键词，提供人们对文本的概要理解。关键词提取技术中基础方法是基于tf-idf，通过词频和到文档频率统计文本中词的tf-idf权重，通过排序得到文本关键词。

基于词图模型的关键词提取方法TextRank。PageRank算法是整个Google搜索的核心算法，是一种通过网页之间的超链接来计算网页重要性的技术，其关键的思想是重要性传递。在关键词提取领域， Mihalcea 等人所提出的TextRank算法就是在文本关键词提取领域借鉴了这种思想。通过设定滑动窗大小，定义词间关联，模拟PageRank的网页关联，实现关键词提取。

其他的方法还有Rake算法和基于主题模型（LDA）的方式实现关键词提取。

#### 全文检索技术

。

#### 文有害文本识别模型技术

文本处理是基于多语种文本索引技术完成，通过对大量的数据进行机器学习，并辅以一定的人工标注，完成对文本的情感倾向，语义解析，行为预测等功能。

#### 图片特征提取技术

#### 图片人像识别技术

图片人像识别技术是基于人像特征信息的生物识别技术，因其非接触式信息采集的优势，被广泛应用于金融、安防、教育、司法、航天等领域。

人像检测是人像识别的基础，人像位置有效检测成为第一个关键步骤，采用基于深度学习的MTCNN（Multi-task convolutional neural network，多任务卷积神经网络），实现人脸区域检测与人脸关键点检测，它的主题框架类似于cascade。MTCNN于2016年中国科学院深圳研究院提出，用于人脸检测任务的多任务神经网络模型，该模型主要采用了三个级联的网络，采用候选框加分类器的思想，进行快速高效的人脸检测。在GPU的支持下可实现实时人像检测。

人像特征提取采用ResNet特征，128维度特征向量。通过大量人的不同人像的测试样本，通过ROC曲线确定人像不同人识别的最佳阈值，实现99.6%的人像识别精度。

#### 图片文字识别技术

图片文字识别分为图片文字定位及语种判别，文字识别两个部分。图片文字定位及语种判别利用深度卷积网络技术，通过海量样本，基于opencv和Tensorflow平台，训练文字定位模型。在深度学习模型的基础上，结合NMS技术对图片中文字进行定位，对无论对图片还是自然场景图片，都具备极高定位精度。同时采用GPU加速技术，具备极高的文字定位效率。

文字识别技术，目前开源技术有基于Leponic的图片处理库的tesseract文字识别技术，采用其训练好的多语言文字识别模型，有效识别图片中定位后文字区域。目前tesseract支持语种达100多种。

#### 分布式集群管控技术

分布式集群主要针对Hadoop生态集群，这里是基于zookeeper完成对Hadoop这个分布式集群的管理控制。它属于Hadoop的正式子项目，但是在kafka等分布式消息中间件当中也会使用。ZooKeeper是一个[分布式](https://baike.baidu.com/item/%E5%88%86%E5%B8%83%E5%BC%8F/19276232" \t "https://baike.baidu.com/item/zookeeper/_blank)的，开放源码的[分布式应用程序](https://baike.baidu.com/item/%E5%88%86%E5%B8%83%E5%BC%8F%E5%BA%94%E7%94%A8%E7%A8%8B%E5%BA%8F/9854429)协调服务，是[Google](https://baike.baidu.com/item/Google)的Chubby一个[开源](https://baike.baidu.com/item/%E5%BC%80%E6%BA%90/246339" \t "https://baike.baidu.com/item/zookeeper/_blank)的实现，是Hadoop和Hbase的重要组件。它是一个为分布式应用提供一致性服务的软件，提供的功能包括：配置维护、域名服务、分布式同步、组服务等。

ZooKeeper的目标就是封装好复杂易出错的关键服务，将简单易用的接口和性能高效、功能稳定的系统提供给用户。

ZooKeeper包含一个简单的原语集， [1]  提供Java和C的接口。

ZooKeeper代码版本中，提供了分布式独享锁、选举、队列的接口，代码在$zookeeper\_home\src\recipes。其中分布锁和队列有[Java](https://baike.baidu.com/item/Java/85979" \t "https://baike.baidu.com/item/zookeeper/_blank)和C两个版本，选举只有Java版本。

#### 分布式应用协调技术

此处的分布式应用协调技术使用的是yarn，同样是属于Hadoop的正式子项目，Apache Hadoop YARN （Yet Another Resource Negotiator，另一种资源协调者）是一种新的 Hadoop 资源管理器，它是一个通用资源管理系统，可为上层应用提供统一的资源管理和调度，它的引入为集群在利用率、资源统一管理和数据共享等方面带来了巨大好处。

#### 离线数据挖掘技术

离线数据挖掘选用hive当做数据仓库，为离线数据提供高效的清理，统计，分析功能。hive是基于Hadoop的一个[数据仓库](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E4%BB%93%E5%BA%93/381916" \t "https://baike.baidu.com/item/hive/_blank)工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的sql查询功能，可以将sql语句转换为MapReduce任务进行运行。 其优点是学习成本低，可以通过类SQL语句快速实现简单的MapReduce统计，不必开发专门的MapReduce应用，十分适合[数据仓库](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E4%BB%93%E5%BA%93/381916" \t "https://baike.baidu.com/item/hive/_blank)的统计分析。

Hive是建立在 Hadoop 上的数据仓库基础构架。它提供了一系列的工具，可以用来进行数据提取转化加载（ETL），这是一种可以存储、查询和分析存储在 Hadoop 中的大规模数据的机制。Hive 定义了简单的类 SQL 查询语言，称为 HQL，它允许熟悉 SQL 的用户查询数据。同时，这个语言也允许熟悉 MapReduce 开发者的开发自定义的 mapper 和 reducer 来处理内建的 mapper 和 reducer 无法完成的复杂的分析工作。

Hive 没有专门的数据格式。 Hive 可以很好的工作在 Thrift 之上，控制分隔符，也允许用户指定数据格式。

#### 分布式离线计算框架

基于spark完成大数据的分析，Apache Spark 是专为大规模数据处理而设计的快速通用的计算引擎。Spark是UC Berkeley AMP lab (加州大学伯克利分校的AMP实验室)所开源的类Hadoop MapReduce的通用并行框架，Spark，拥有Hadoop MapReduce所具有的优点；但不同于MapReduce的是——Job中间输出结果可以保存在内存中，从而不再需要读写HDFS，因此Spark能更好地适用于数据挖掘与机器学习等需要迭代的MapReduce的算法。

Spark 是一种与 Hadoop 相似的开源集群计算环境，但是两者之间还存在一些不同之处，这些有用的不同之处使 Spark 在某些工作负载方面表现得更加优越，换句话说，Spark 启用了内存分布数据集，除了能够提供交互式查询外，它还可以优化迭代工作负载。

Spark 是在 [Scala](https://baike.baidu.com/item/Scala/2462287) 语言中实现的，它将 Scala 用作其应用程序框架。与 Hadoop 不同，Spark 和 [Scala](https://baike.baidu.com/item/Scala/2462287) 能够紧密集成，其中的 [Scala](https://baike.baidu.com/item/Scala/2462287) 可以像操作本地集合对象一样轻松地操作分布式数据集。

尽管创建 Spark 是为了支持分布式数据集上的迭代作业，但是实际上它是对 Hadoop 的补充，可以在 Hadoop 文件系统中并行运行。通过名为 Mesos 的第三方集群框架可以支持此行为。Spark 由加州大学伯克利分校 AMP 实验室 (Algorithms, Machines, and People Lab) 开发，可用来构建大型的、低延迟的数据分析应用程序。

#### 分布式实时流计算技术

Spark Streaming：构建在Spark上处理Stream数据的框架，基本的原理是将Stream数据分成小的时间片段（几秒），以类似batch批量处理的方式来处理这小部分数据。Spark Streaming构建在Spark上，一方面是因为Spark的低延迟执行引擎（100ms+），虽然比不上专门的流式数据处理软件，也可以用于实时计算，另一方面相比基于Record的其它处理框架（如Storm），一部分窄依赖的RDD数据集可以从源数据重新计算达到容错处理目的。此外小批量处理的方式使得它可以同时兼容批量和实时数据处理的逻辑和算法。方便了一些需要历史数据和实时数据联合分析的特定应用场合。

#### 机器学习技术

基于TensorFlow完成机器学习，TensorFlow™是一个基于[数据流编程](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%B5%81%E7%BC%96%E7%A8%8B/22735640" \t "https://baike.baidu.com/item/TensorFlow/_blank)（dataflow programming）的符号数学系统，被广泛应用于各类[机器学习](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/217599" \t "https://baike.baidu.com/item/TensorFlow/_blank)（machine learning）算法的编程实现，其前身是[谷歌](https://baike.baidu.com/item/%E8%B0%B7%E6%AD%8C/117920" \t "https://baike.baidu.com/item/TensorFlow/_blank)的神经网络算法库DistBelief。

Tensorflow拥有多层级结构，可部署于各类[服务器](https://baike.baidu.com/item/%E6%9C%8D%E5%8A%A1%E5%99%A8/100571" \t "https://baike.baidu.com/item/TensorFlow/_blank)、PC终端和[网页](https://baike.baidu.com/item/%E7%BD%91%E9%A1%B5/99347" \t "https://baike.baidu.com/item/TensorFlow/_blank)并支持[GPU](https://baike.baidu.com/item/GPU/105524)和[TPU](https://baike.baidu.com/item/TPU/20473545" \t "https://baike.baidu.com/item/TensorFlow/_blank)高性能[数值计算](https://baike.baidu.com/item/%E6%95%B0%E5%80%BC%E8%AE%A1%E7%AE%97/3729797" \t "https://baike.baidu.com/item/TensorFlow/_blank)，被广泛应用于谷歌内部的产品开发和各领域的科学研究。

TensorFlow由谷歌[人工智能](https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD/9180" \t "https://baike.baidu.com/item/TensorFlow/_blank)团队[谷歌大脑](https://baike.baidu.com/item/%E8%B0%B7%E6%AD%8C%E5%A4%A7%E8%84%91/4649855)（Google Brain）开发和维护，拥有包括TensorFlow Hub、TensorFlow Lite、TensorFlow Research Cloud在内的多个项目以及各类[应用程序接口](https://baike.baidu.com/item/%E5%BA%94%E7%94%A8%E7%A8%8B%E5%BA%8F%E6%8E%A5%E5%8F%A3/10418844" \t "https://baike.baidu.com/item/TensorFlow/_blank)（Application Programming Interface, API）。自2015年11月9日起，TensorFlow依据[阿帕奇授权协议](https://baike.baidu.com/item/%E9%98%BF%E5%B8%95%E5%A5%87%E6%8E%88%E6%9D%83%E5%8D%8F%E8%AE%AE/1642155" \t "https://baike.baidu.com/item/TensorFlow/_blank)（Apache 2.0 open source license）[开放源代码](https://baike.baidu.com/item/%E5%BC%80%E6%94%BE%E6%BA%90%E4%BB%A3%E7%A0%81/114160)。

#### 分布式文件系统技术

基于fastdfs的轻量级小文件存储系统。FastDFS是一个开源的轻量级[分布式文件系统](https://baike.baidu.com/item/%E5%88%86%E5%B8%83%E5%BC%8F%E6%96%87%E4%BB%B6%E7%B3%BB%E7%BB%9F/1250388" \t "_blank)，它对文件进行管理，功能包括：文件存储、文件同步、文件访问（文件上传、文件下载）等，解决了大容量存储和负载均衡的问题。特别适合以文件为载体的在线服务，如相册网站、视频网站等等。

FastDFS为互联网量身定制，充分考虑了冗余备份、负载均衡、线性扩容等机制，并注重高可用、高性能等指标，使用FastDFS很容易搭建一套高性能的文件服务器集群提供文件上传、下载等服务。

#### 列式数据库技术

基于hbase完成数据存储，借助列式数据库的特性，可以根据业务具体需要新增或删减列，以key-value的形式进行数据查询，达到高效快速。HBase是一个分布式的、面向列的开源数据库，该技术来源于 Fay Chang 所撰写的Google论文“Bigtable：一个结构化数据的[分布式存储系统](https://baike.baidu.com/item/%E5%88%86%E5%B8%83%E5%BC%8F%E5%AD%98%E5%82%A8%E7%B3%BB%E7%BB%9F/6608875" \t "https://baike.baidu.com/item/_blank)”。就像Bigtable利用了Google文件系统（File System）所提供的分布式数据存储一样，HBase在Hadoop之上提供了类似于Bigtable的能力。HBase是Apache的Hadoop项目的子项目。HBase不同于一般的关系数据库，它是一个适合于非结构化数据存储的数据库。另一个不同的是HBase基于列的而不是基于行的模式。

#### 文档性数据库技术

在面对文档型数据时，我们选择MongoDB作为存储数据库。MongoDB是一个基于分布式文件存储的数据库。由[C++](https://baike.baidu.com/item/C%2B%2B" \t "_blank)语言编写。旨在为WEB应用提供可扩展的高性能数据存储解决方案。

MongoDB是一个介于[关系数据库](https://baike.baidu.com/item/%E5%85%B3%E7%B3%BB%E6%95%B0%E6%8D%AE%E5%BA%93" \t "_blank)和非关系数据库之间的产品，是非关系数据库当中功能最丰富，最像关系数据库的。它支持的数据结构非常松散，是类似[json](https://baike.baidu.com/item/json)的[bson](https://baike.baidu.com/item/bson" \t "_blank)格式，因此可以存储比较复杂的数据类型。Mongo最大的特点是它支持的查询语言非常强大，其语法有点类似于面向对象的查询语言，几乎可以实现类似关系数据库单表查询的绝大部分功能，而且还支持对数据建立[索引](https://baike.baidu.com/item/%E7%B4%A2%E5%BC%95" \t "_blank)。

#### 内存数据库技术

内存数据库有memcache，redis，redis-cluster等，由于业务特点和数据使用情况，选择redis和redis-cluster当做内存数据库。Redis（全称：Remote Dictionary Server 远程字典服务）是一个开源的使用ANSI [C语言](https://baike.baidu.com/item/C%E8%AF%AD%E8%A8%80)编写、支持网络、可基于内存亦可持久化的日志型、Key-Value[数据库](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BA%93/103728)，并提供多种语言的API。从2010年3月15日起，Redis的开发工作由VMware主持。从2013年5月开始，Redis的开发由Pivotal赞助。

redis是一个key-value[存储系统](https://baike.baidu.com/item/%E5%AD%98%E5%82%A8%E7%B3%BB%E7%BB%9F)。和Memcached类似，它支持存储的value类型相对更多，包括string(字符串)、list([链表](https://baike.baidu.com/item/%E9%93%BE%E8%A1%A8))、set(集合)、zset(sorted set --有序集合)和hash（哈希类型）。这些[数据类型](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E7%B1%BB%E5%9E%8B" \t "https://baike.baidu.com/item/_blank)都支持push/pop、add/remove及取交集并集和差集及更丰富的操作，而且这些操作都是原子性的。在此基础上，redis支持各种不同方式的排序。与memcached一样，为了保证效率，数据都是缓存在内存中。区别的是redis会周期性的把更新的数据写入磁盘或者把修改操作写入追加的记录文件，并且在此基础上实现了master-slave(主从)同步。

Redis 是一个高性能的key-value数据库。 redis的出现，很大程度补偿了[memcached](https://baike.baidu.com/item/memcached" \t "https://baike.baidu.com/item/_blank)这类key/value存储的不足，在部 分场合可以对关系数据库起到很好的补充作用。它提供了Java，C/C++，C#，PHP，JavaScript，Perl，Object-C，Python，Ruby，Erlang等客户端，使用很方便。

#### 关系型数据库技术

基于MySQL、mysql\_galera项目进行数据的管理。MySQL是一个关系型数据库管理系统，在 WEB 应用方面，MySQL是RDBMS （Relational Database Management System，关系数据库管理系统）应用软件。MySQL是一种关联数据库管理系统，关联数据库将数据保存在不同的表中，而不是将所有数据放在一个大仓库内，这样就增加了速度并提高了灵活性。MySQL所使用的 SQL 语言是用于访问数据库的最常用标准化语言。MySQL 软件具有体积小、速度快、总体拥有成本低等特点。

Galera是一个MySQL(也支持MariaDB，Percona)的同步多主集群软件。具有同步复制、Active-active的多主拓扑结构、真正的multi-master、自动成员资格控制，失败节点从群集中删除、新节点加入数据自动复制、真正的并行复制，行级的功能，在数据备份方面有着很大的优势

#### 图数据库技术

基于neo4j完成图数据的存储，查询，达到知识库的积累。[Neo4j](https://baike.baidu.com/item/Neo4j" \t "_blank)是一个高性能的,NOSQL图形数据库，它将结构化数据存储在网络上而不是表中。它是一个[嵌入式](https://baike.baidu.com/item/%E5%B5%8C%E5%85%A5%E5%BC%8F/575465" \t "_blank)的、基于[磁盘](https://baike.baidu.com/item/%E7%A3%81%E7%9B%98/2842227)的、具备完全的事务特性的Java持久化引擎，但是它将结构化数据存储在网络(从数学角度叫做图)上而不是表中。Neo4j也可以被看作是一个高性能的图引擎，该引擎具有成熟数据库的所有特性。程序员工作在一个面向对象的、灵活的网络结构下而不是严格、静态的表中——但是他们可以享受到具备完全的事务特性、企业级的数据库的所有好处。

Neo4j因其嵌入式、高性能、轻量级等优势，越来越受到关注.

#### 时序数据库技术

。

#### 网络数据采集技术

数据采集使用Python语言完成spider任务，Python是一种跨平台的[计算机程序设计语言](https://baike.baidu.com/item/%E8%AE%A1%E7%AE%97%E6%9C%BA%E7%A8%8B%E5%BA%8F%E8%AE%BE%E8%AE%A1%E8%AF%AD%E8%A8%80/7073760" \t "https://baike.baidu.com/item/Python/_blank)。是一种面向对象的动态类型语言，最初被设计用于编写自动化脚本(shell)，随着版本的不断更新和语言新功能的添加，越来越多被用于独立的、大型项目的开发。Python已经成为最受欢迎的[程序设计语言](https://baike.baidu.com/item/%E7%A8%8B%E5%BA%8F%E8%AE%BE%E8%AE%A1%E8%AF%AD%E8%A8%80/2317999" \t "https://baike.baidu.com/item/Python/_blank)之一。自从2004年以后，python的使用率呈线性增长。Python 2于2000年10月16日发布，稳定版本是Python 2.7。Python 3于2008年12月3日发布，不完全兼容Python 2。2011年1月，它被[TIOBE](https://baike.baidu.com/item/TIOBE" \t "https://baike.baidu.com/item/Python/_blank)编程语言[排行榜](https://baike.baidu.com/item/%E6%8E%92%E8%A1%8C%E6%A6%9C/4895" \t "https://baike.baidu.com/item/Python/_blank)评为2010年度语言。

#### 数据迁移技术

。

#### 海量日志采集技术

使用filebeat+logstash完成海量日志采集。Filebeat是本地文件的日志数据采集器。 作为服务器上的代理安装，Filebeat监视日志目录或特定日志文件，tail file，并将它们转发给Elasticsearch或Logstash进行索引、kafka 等。Logstash是一个具有实时管道功能的开源数据收集引擎，Logstash可以动态地将来自不同数据源的数据统一起来，并将数据规范化为你选择的目的地，清理和大众化你的所有数据，用于各种高级下游分析和可视化用例。虽然Logstash最初推动了日志收集方面的创新，但是它的功能远远超出了这个用例，任何类型的事件都可以通过大量的输入、过滤器和输出插件来丰富和转换，使用许多原生编解码可以进一步简化摄取过程。Logstash通过利用大量和多种数据来提高你的洞察力。

#### 分布式消息订阅处理技术

消息订阅基于kafka消息中间件，Kafka [1]  是一种高吞吐量的分布式发布订阅消息系统，有如下特性：通过O(1)的磁盘数据结构提供消息的持久化，这种结构对于即使数以TB的消息存储也能够保持长时间的稳定性能。高吞吐量 ：即使是非常普通的硬件Kafka也可以支持每秒数百万的消息。支持通过Kafka服务器和消费机集群来分区消息。支持[Hadoop](https://baike.baidu.com/item/Hadoop)并行数据加载。

#### ETL技术

。

#### 数据脱敏技术

。

#### 数据交换技术

。

#### 数据仓库技术

。

#### 数据环境监控技术

。

#### 数据生命周期技术

。