# 姨搜和知识图谱

## —如何利用大数据来做自动化风控

侯松
songhou@creditease.cn

# Table of Contents

# 1. 我们是谁

- 宜信：世界上规模最大的P2P金融公司之一。
- 姨搜团队：为整个宜信公司提供风控数据服务和模型服务，包括风控搜索引擎，大规模知识图谱以及其上的图谱搜索等重量级产品。
- 本人：大数据系统的研发有超过5年的实战经验，包括且不限于Hadoop、Hbase、Hive、Spark、ES、大规模语义网和查询引擎、实时流式计算、工作流调度执行引擎、监控报警系统等等。

# 2. 我们做了什么

1. 风控搜索引擎

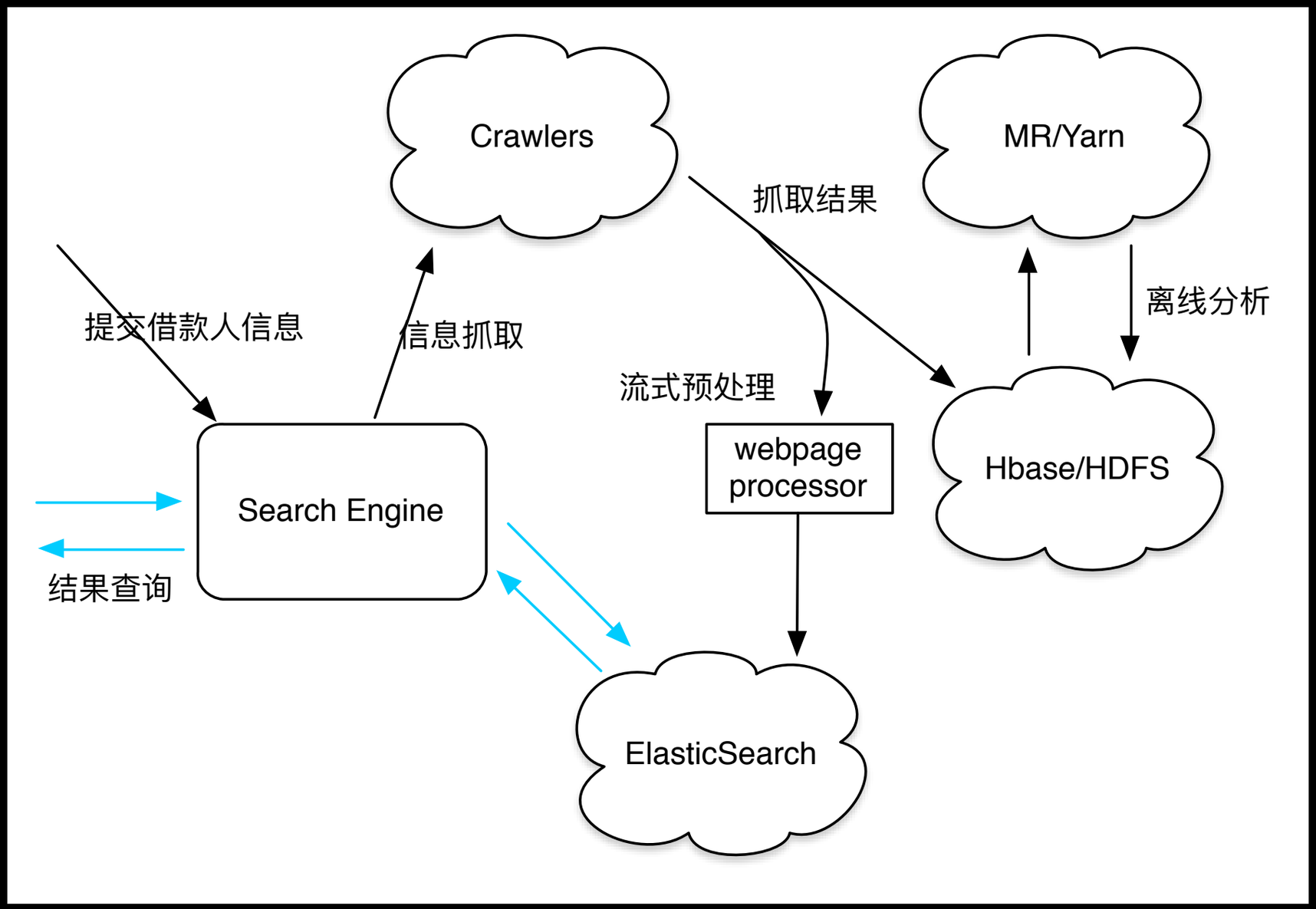
2. 大规模知识图谱

3. 图谱搜索

# 2.1 风控搜索引擎

- 利用爬虫抓取借款人在网上的各种信息。
- 利用机器学习得到的一些规则来提取重要信息，过滤无效数据。
- 对重要信息重排序，将更可信、更相关的条目排到前面。
- 融合进信审的标准流程，由审核人员进行判断。

# 2.1 风控搜索引擎


Crawlers
MR/Yarn
抓取结果
提交借款人信息
信息抓取
离线分析
流式预处理
webpage
processor
Hbase/HDFS
Search Engine
结果查询
ElasticSearch

# 2.2 大规模知识图谱

- 风控的核心能力，是能从纷繁复杂的数据中辨识出和用户相关的信息，并能进行交叉验证和风险程度评估。

- 知识图谱利用传统语义网的部分原理，将各个数据源不同格式和类别的数据融合到同一个可扩展的平台下，用相同的流程同时分析所有的数据。

- 构建于Hadoop、Hbase、ES等的开源环境之上，拥有几乎线性的扩展能力，并且提供一系列的工具链来帮助开发人员和最终用户来轻松驾驭这些数据。

# 2.2 大规模知识图谱

1. 知识图谱的基本概念
2. 知识图谱体系的组成
3. 如何实现高度可扩展的知识图谱
4. 如何使用知识图谱带来的能力

# 2.2.1 知识图谱的概念

- 数据的表现形式，可以是文本文件，可以是定义好schema的数据库表，可以是NoSQL中的键值对，也可以是自然知识中的entity-property-link的形式。
- 知识图谱使用最后一种形式，拥有强大的表现能力和扩展能力，为knowledge-based reasoning提供基础。
    - 以entity为中心，每个entity拥有独特的属性值，entity之间有多样的关联关系。
    - 灵活的schema，增加新的知识更加简单。
    - 新应用可以无缝重用已有数据，加快开发进度。

# 2.2.2 知识图谱体系的组成

1. 元数据的定义和描述

2. 数据的获取和转化

3. 数据间关系的确立

4. 提高系统的智能

5. 简单易用的工具

# 2.2.2 知识图谱体系的组成

1. 元数据的定义和描述

   A. 元数据的定义需要简单易用，无歧义，扩展能力强。我们使用了完整的RDF schema，部分的OWL，以及部分的自定义功能，比如数据格式的验证规则等。

   B. 建立元数据委员会，负责评审，确保正确性。

   C. 使用自动化的方式，从已有的数据源导入数据，比如freebase，dbpedia等。

# 2.2.2 知识图谱体系的组成

2. 数据的获取和转化

A. 获取方式：可以是被动的接收数据，也可以是主动的通过爬虫或者DB查询来获得数据。

B. 数据格式：可能是HTML网页、excel文档、数据库表，甚至可能是纯文本。

C. 时效性：有些要求实时入库实时查询，有些则T+1以上也可以接受。

# 2.2.2 知识图谱体系的组成

3. 数据间关系的确立

A. 显式关系，是输入数据内含的关系，很多的逻辑演绎和推论都建立在这些关系的基础上。

B. 隐式关系，是通过特征匹配来查找相关的entity。隐式关系的确立更加复杂和多样，同时也是KG发挥最大用途的基础。

# 2.2.2 知识图谱体系的组成

4. 提高系统的智能

A. 知识图谱的schema给数据带来了丰富的语义，根据这些语义进行的推理是提升系统智能的关键。

B. 知识图谱体系应该提供易用可扩展的推理能力，为构建更复杂的推理逻辑提供基础。

# 2.2.2 知识图谱体系的组成

5. 简单易用的工具

A. 和不同类型的数据源同步，数据交换的需求。

B. 灵活多样的API接口，满足不同类型的开发需要。

C. 支持强大语法的查询引擎，能将复杂业务逻辑（OLAP/OLTP）解析并执行出结果。

# 2.2.2 知识图谱体系的组成

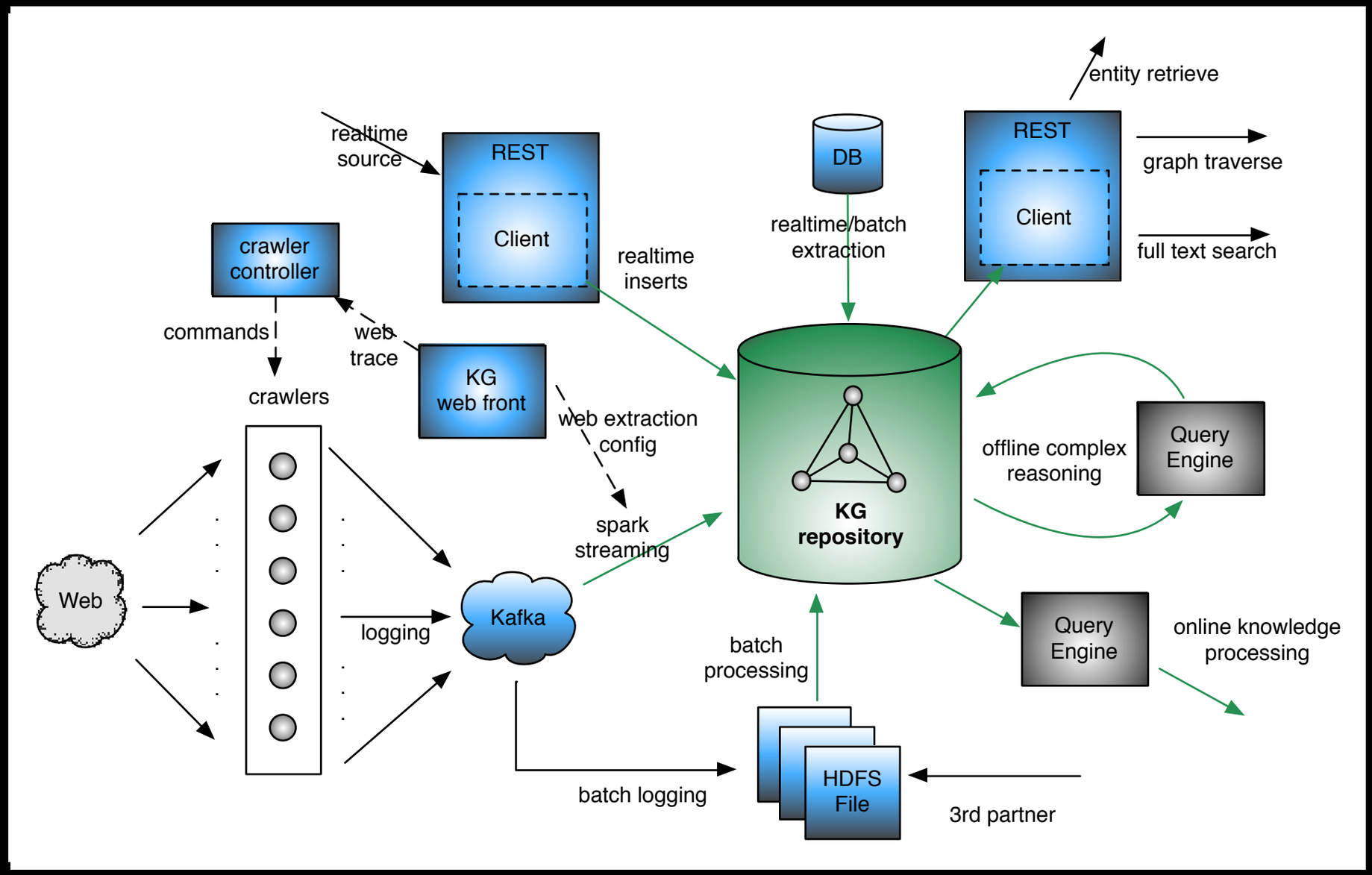
```
PREFIX  rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX  rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  e:    <http://bdp.creditease.cn/rdf/entity/>
PREFIX  m:    <http://bdp.creditease.cn/rdf/schema/>

SELECT  ?person (COUNT(?black_entity) as ?blacklist_addr_count)
FROM e:
FROM NAMED m:
WHERE
  {
    ?person m:finance.mortgagor.apply_id "12345" .
    ?person m:people.person.sfz_id ?sfz .
    ?person m:people.person.company/m:org.org.addr/m:common.location.detail.address ?company_addr .
    ?black_entity m:blacklist.entity.address ?black_addr
    FILTER (?black_addr = ?company_addr)
  }
  GROUP BY ?person
```

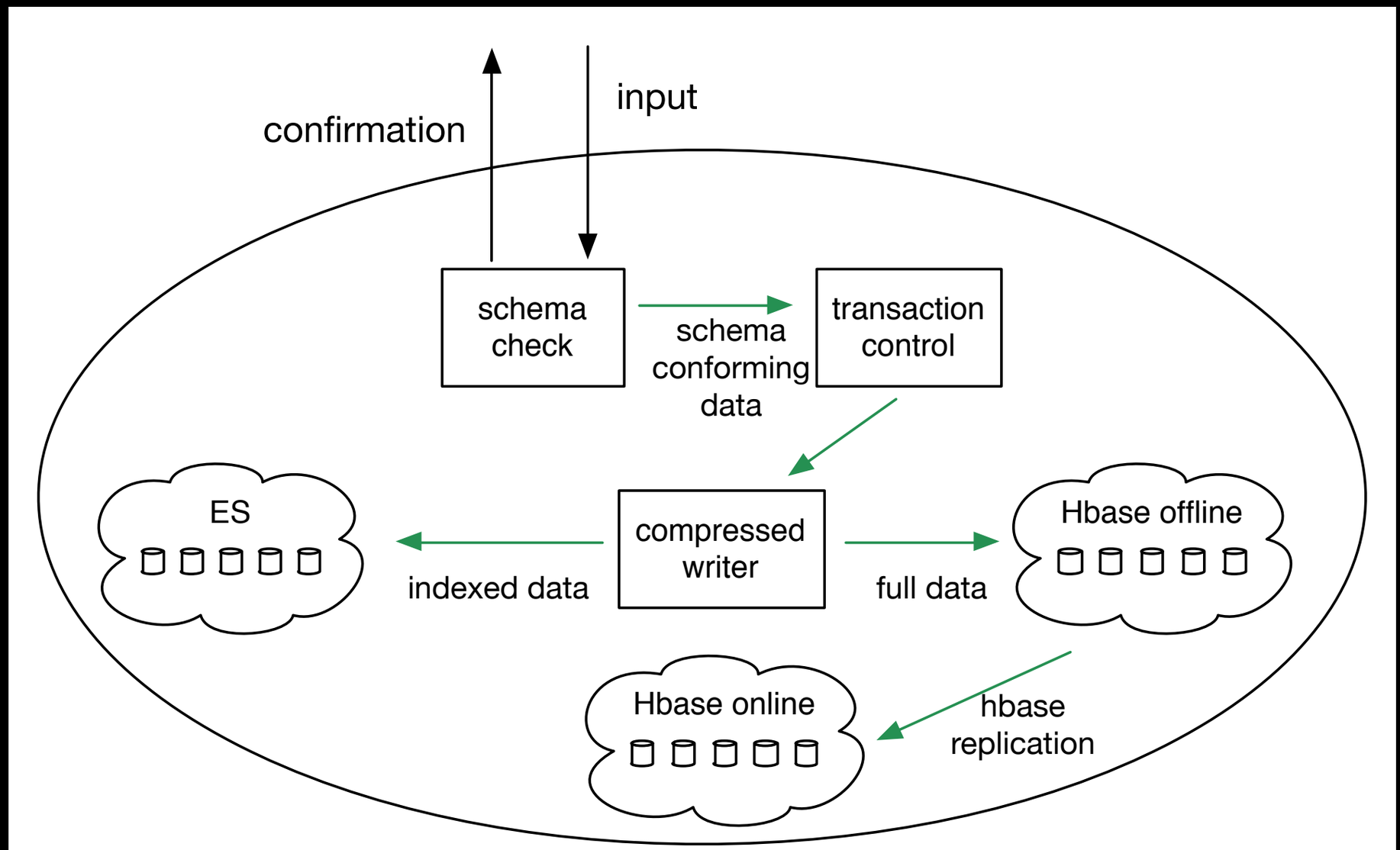
# 2.2.3 如何实现高度可扩展的知识图谱

1. 知识图谱的图的本质，要求有一个同时支持离线分析和在线服务的大规模图数据库。
2. 图里面的知识内容有很多是非结构化信息，需要有全文检索能力。
3. 部分知识的时效性很强，需要支持实时写入。
4. 支持透明压缩，节省计算资源，提升整体性能。

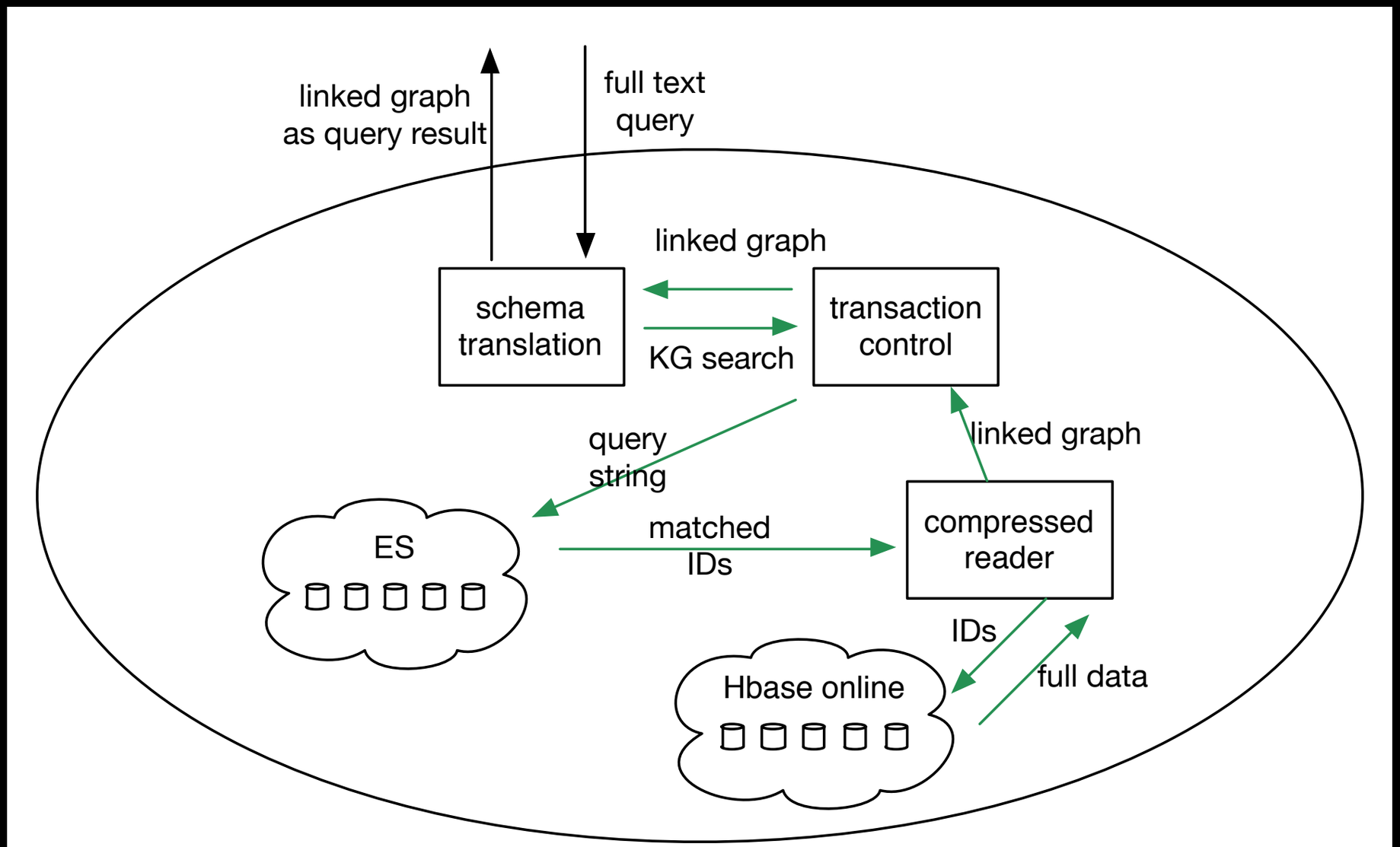
# 2.2.3 如何实现高度可扩展的知识图谱



# 2.2.3 如何实现高度可扩展的知识图谱



# 2.2.3 如何实现高度可扩展的知识图谱



# 2.2.4 如何使用知识图谱带来的能力

1. 无缝融合任意数据的能力：将组织的内部外部数据全部整合到一起，统一管理统一分析。
2. 丰富的语义信息：知识图谱的所有数据都带有语义信息，数据分析人员更容易理解和开展分析工作。
3. 网状结构的连接数据：轻易的看到数据关联，自动的隐含关联关系发现的能力进一步提高数据处理的自动化水平。
4. 易用的查询引擎：给非技术人员提供了使用知识图谱的能力，即支持毫秒级在线即时查询也支持大规模离线分析。

# 2.2 大规模知识图谱

1. 知识图谱的基本概念
2. 知识图谱体系的组成
3. 如何实现高度可扩展的知识图谱
4. 如何使用知识图谱带来的能力

# 2.3 图谱搜索

1. 基于知识图谱开发的线索分析平台。

2. 广泛用于欺诈案件的识别和规则发现。

# 2.3 图谱搜索

# 2. 我们做了什么

1. 风控搜索引擎
2. 大规模知识图谱
3. 图谱搜索

The open source Palantir

# 3. 有什么样的收获

1. 大数据风控不再是一个虚幻的概念，我们可以做的事情有很多。
2. 研发人员需要既有广泛的涉猎又有深入的钻研，集百家之所长，才能创造性的解决现实问题并预见性的提出新问题。
3. 数据融合和使用变得更加容易，使得数据所有者有了更大的话语权。

# Thanks !
# Q & A

侯松
songhou@creditease.cn
http://housong.github.io/