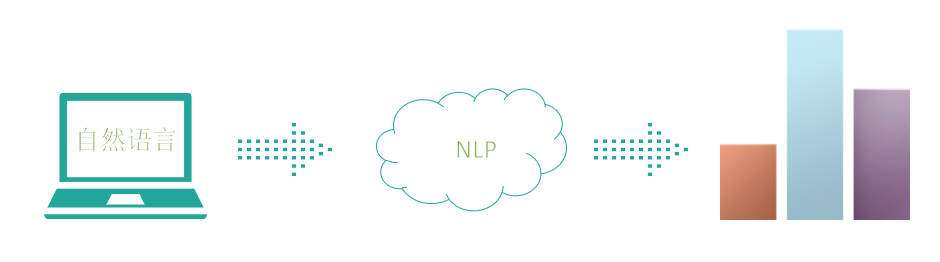
**构建智能BI系统 理解报告**

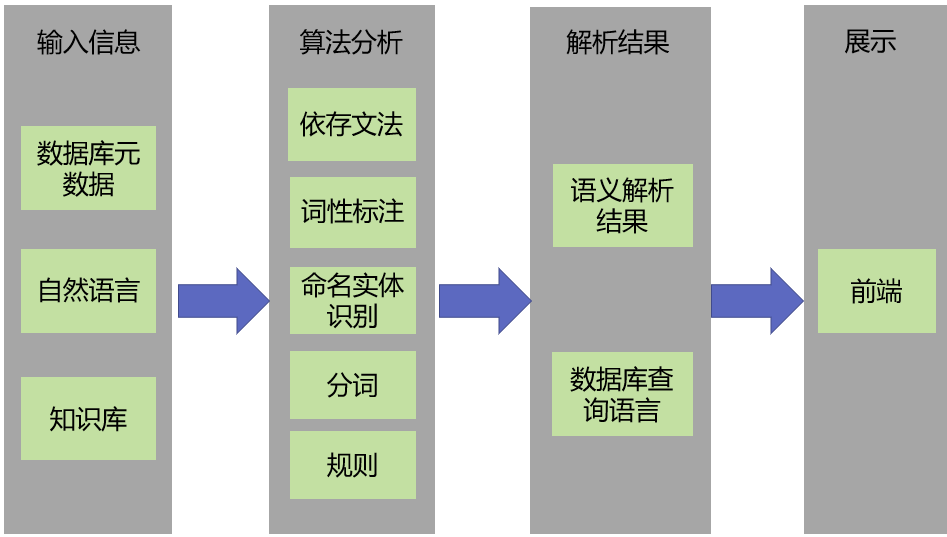
**智能BI 系统**

BI 系统为商务活动提供数据支撑，用户可以根据业务的需求查询相关数据，也可以获取数据的可视化图表。通常来说业务人员需要在BI系统中进行频繁地拖拽操作，专业人员也可以直接使用SQL 语句进行相关操作，但是这样操作会比较麻烦。智能是指系统能够更加深层次地理解用户的意图。用户输入一条自然语言，系统可以根据用户的输入准确输出用户需要的图表。智能BI系统可以降低业务人员的使用门槛，同时也大大降低了操作复杂度。



**项目流程理解**

数据库元数据、用户输入的自然语言和外部知识库作为输入，经过NLP算法处理，解析为格式工整的json数据，最终转换为数据库查询语言。



项目中核心为自然语言处理技术，主要有分词，词性标注，实体识别，依存句法分析，定义解析规则。

1. 分词

分词往往是文本处理的首要任务。而中文分词相对英文分词会困难很多，比如“北京大学生”这句话如果按照词汇表进行分词，就有[“北京大学”，“生”]和

[“北京”，“大学生”]两种不同的分词法，而两种分词方法造就的意思则差别很大。现代中文分词常常基于统计学，机器学习等方法，分词的准确率也大大提高。具体在项目中可以使用结巴分词、哈工大LTP等已有工具进行分词。

1. 词性标注

词性标注具有比较浓重的语言学色彩，是给句子分词以后的每一个词标注词性类别，有名词，副词，动词等。一些成熟的工具词性标注的粒度会很细，如名词可以分为位置名词、地理名词、时间名词等。词性标注是后续实体识别、句法分析等工作的基础。

1. 命名实体识别

命名实体识别的任务是在句子的词序列中定位并识别出人名、地名、机构名等实体名称。例如：

国家主席（机构名）习近平（人名）在上海（地名）调研。

挖掘文本中的实体，并且针对实体进行分析非常关键。

1. 依存句法分析

通过分析句子之间的依存关系从而揭示其语法结构。依存句法分析识别出句子中的“主谓宾”、“定状补”这些语法成分，并分析各成分之间的关系。

1. 定义解析规则

用户的自然语言经过上述步骤的处理之后就会变成结构化的数据（可以是json, xml 等），这个时候就可以针对这些结构化的数据，定义解析规则。根据业务的需求，将结构化的数据转化为数据库查询语言。如果业务复杂，能力足够，则解析的规则可以使用算法，如果业务需求比较简单，也可以自己定义解析规则。

**项目开发过程需要考虑的问题**

1. 需要调研合法的自然语言输入 自然语言的输入组合不计其数，即使是资深的语言学家，也难以保证可以理解任何形式的语言（比如非常生僻的文言文表达）。所以有必要定义合法的自然语言输入。可以在自然语言与SQL关键字中完成映射的工作。比如SUM是SQL聚合函数之一，而SUM对应的自然语言输出可以是“之和”、“总额”、“总量”等。
2. 定义解析规则时需要考虑的问题：选择合适的图表；当用户的查询输入中包括“占比”时，则使用饼状图更好，包含“趋势”时则使用折线图更好；
3. 合理处理语意缺失、包含隐藏语意、包含非法输入的状况。比如用户输入“10月份销售额”，那么有用户的意思可能是今年10月，也有可能是去年10月
4. 定义需要的外部知识库。比如用户如果输入“上个月的销售额”那么系统必须包含时间库才可以准确识别用户需要查询的时间。“华东地区上半年的销售量之和”，则系统也必须知道华东地区究竟包含哪些省份。
5. 自然语言识别的实体通常对应到数据库中相应的表和字段 。需要根据分析的结果查询表中的数据，实际的项目中，需要考虑的问题有如何映射？如果自然语言中识别的字段找不到怎么办？找到多个如何选择？识别的字段和数据库中的字段不一致等，都是在具体项目中需要考虑的问题。

**项目准备**

1. 知识学习

自然语言处理：分词，词性标注，实体识别，句法分析,词向量等。

语言学：主谓宾、定状补等

可能会用到的：统计学习，深度学习等

2. 工具学习

可能会用到的工具：自然语言处理：Stanford NLP, 哈工大LTP, jieba, Gensim,复旦NLP等

Python: pyCharm, wingIDE, python json工具等。

3. 文献学习及经验学习

参考前人的工作，百分点已经完成的工作，开源的英语转sql工具等