大数据背景下电网数据质量研究与实践

全球能源互联网研究院

2017年8月

国家电网
STATE GRID

全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

目 录

全球能源互联网与大数据

大数据质量提升的系统性方案

大数据清洗治理系统

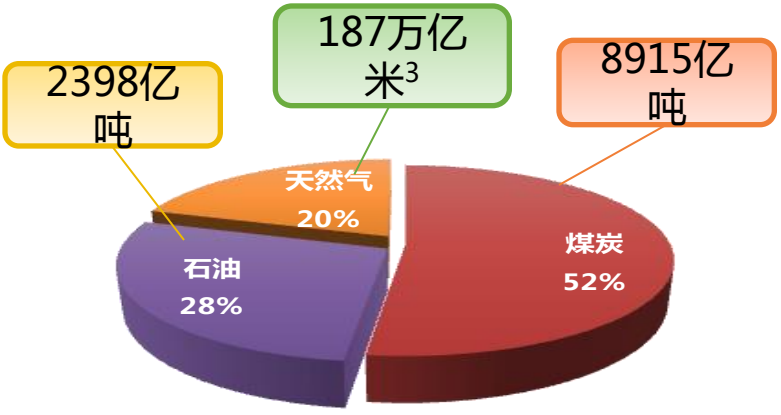
国家电网
STATE GRID
全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

人类可持续发展面临巨大的能源挑战

随着社会经济的发展，人们对于能源的需求越来越大，能源是现代工业社会的命脉，能源问题关乎生态环境和可持续发展。化石能源的大规模开发使用导致我们被迫面临三大严峻挑战，给人类生存发展带来严重威胁。

1 资源紧张

- 煤炭：只能开采110年
- 石油：只能开采57年
- 天然气：只能开采54年



全球化石能源剩余探明可采储量

2 环境污染

化石能源的大量开发，在生产、运输、存储、使用的各环节，对大气、水质、土壤、地貌等造成严重污染和破坏。

3 气候变暖

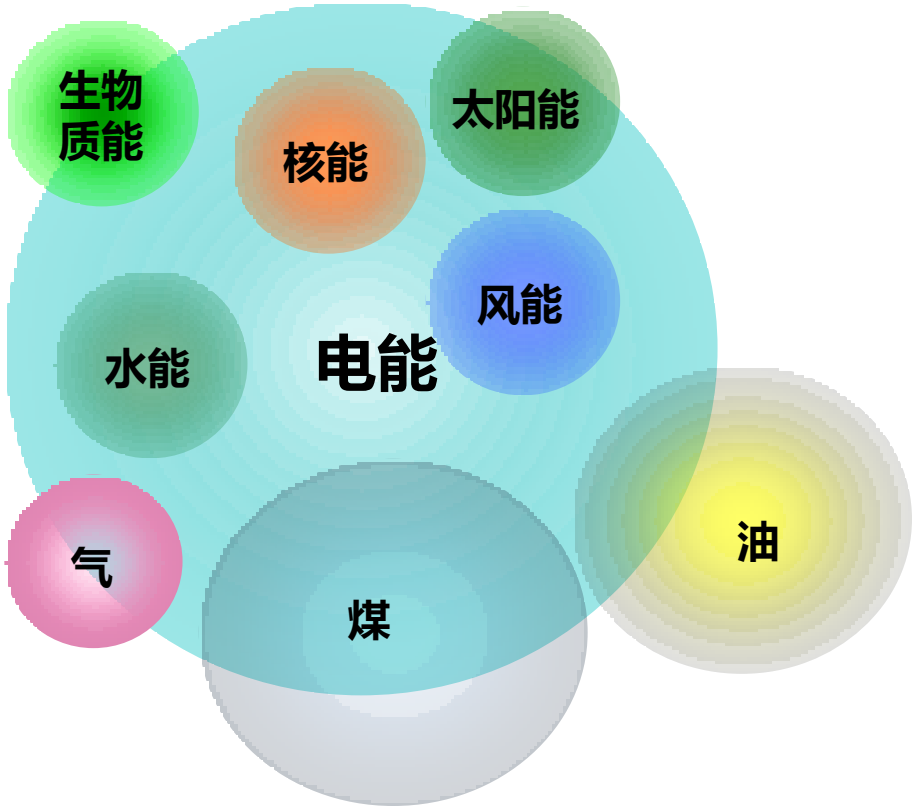
化石能源的碳排放是气候变暖的主因。自1850年以来，全球地表平均温升已经超过1℃。

开发清洁能源体系满足人类可持续发展

清洁能源 - 电能 - 化石能源

- 能源供应侧实施清洁替代：清洁能源替代化石能源
- 能源消费侧实施电能替代：电能替代煤、油、气。

形成电为中心的能源格局



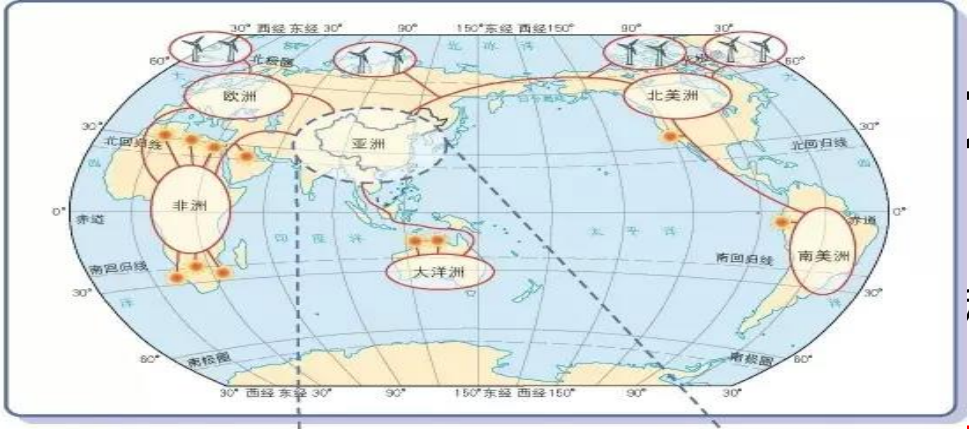
国际能源署（IEA）《2014能源技术展望》（ETP2014）：
推进电气化是全球能源系统的驱动力之一，可从根本上转变能源供应及终端用能

国家电网
STATE GRID

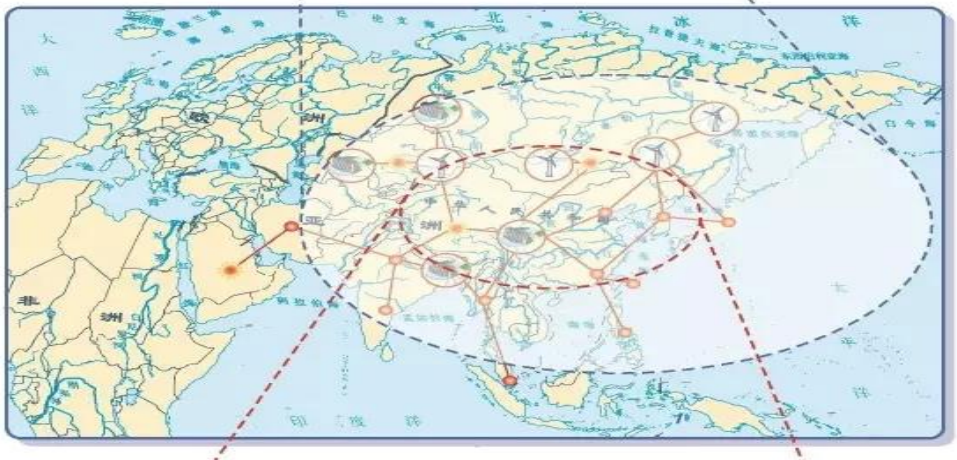
全球能源互
GLOBAL ENERGY INTER

持续发展的必由之路

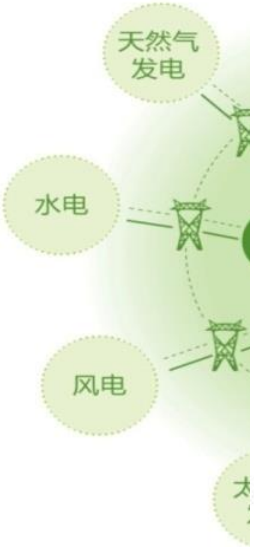
构建以 和绿色方式满足全球电
力需求，是
电网必 网。

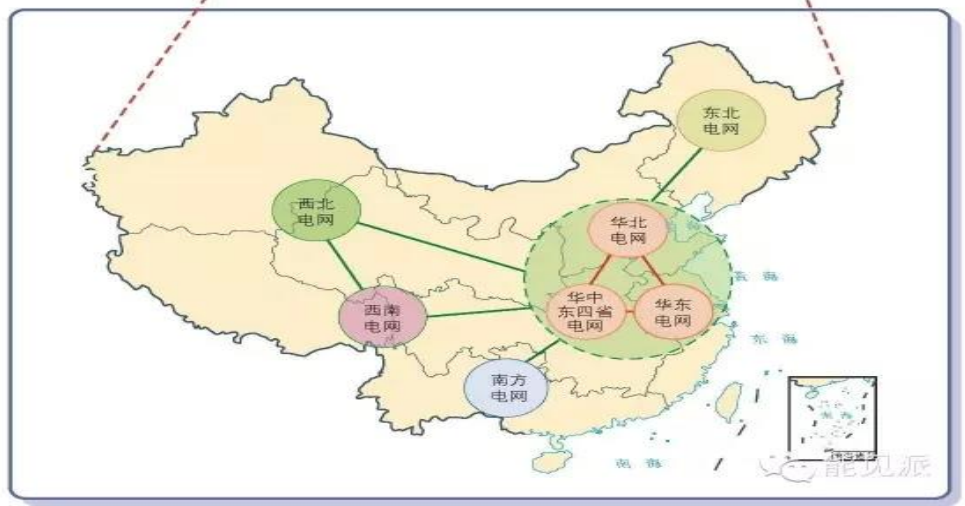


洲际互联

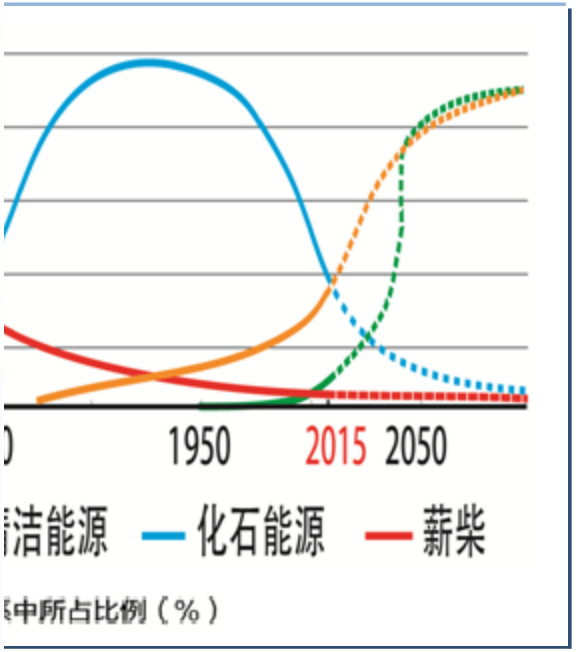


洲内互联





国内互联



电网配置 **低碳发展**

国家电网
STATE GRID
全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

全球能源互联网实质及关键技术

全球能源互联网：
以特高压电网为骨干网架，以输送清洁能源为主导，全球互联泛在的智能电网。
是清洁能源在全球范围大规模开发、配置、利用的基础平台。

特征：网架坚强、广泛互联、高度智能、开放互动

关键技术

- 电源技术：提高清洁能源开发效率和经济性
  - 风力发电、太阳能发电、海洋能发电及分布式电源
- 电网技术：提高输电距离和容量，能源资源全球配置
  - 特高压、海底电缆、超导输电、直流电网、微电网
- 储能技术：保障清洁能源大规模发展和电网安全运行
- **信息通信技术：实现电网智能化、互动化和运行控制**

国家电网
STATE GRID
全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

先进信息技术 - 让电网更加智能

智能电网广泛采用先进信息技术：

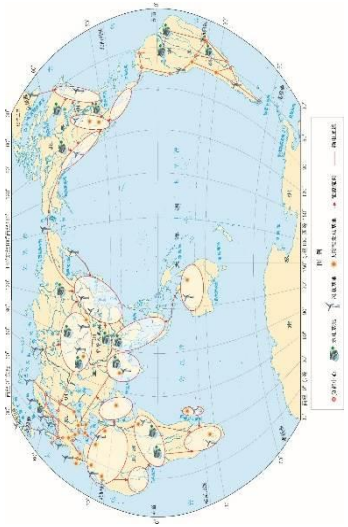
- 自动预判、识别各类电网故障和风险;
- 适应各类集中式、分布式清洁能源大规模接入和大范围配置要求；
- 满足用户多样化、智能化用电需求，构建公共服务平台，促进智能家居、智能社区、智能交通、智慧城市发展。

- 2016年2月29日，国家发改委、能源局以及工信部联合下发《关于推进“互联网+”智慧能源发展的指导意见》（下简称《指导意见》），以促进能源和信息深度融合，推动能源互联网新技术、新模式和新业态发展，推动能源领域供给侧结构性改革和能源革命。

国家电网
STATE GRID

全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

信息技术创新应用 - 大数据



全球能源互联网

大数据是全球能源互联网的必然产物

大数据产品是全球能源互联网的重要服务形态

大数据是支撑全球互联网运行控制的必要技术

大数据

- 知识发现与价值挖掘
- 跨专业、跨领域整合分析
- 海量数据处理能力

国家电网
STATE GRID
全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

大数据技术应用困局

大数据技术应用远比想象的困难！

大数据应用

1 数据质量问题

2 缺少平台级的智能工具

如何唤醒沉睡的文本、视频、图像数据

3 计算的性能难以达到预期

国家电网
STATE GRID

全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

目 录

全球能源互联网与大数据

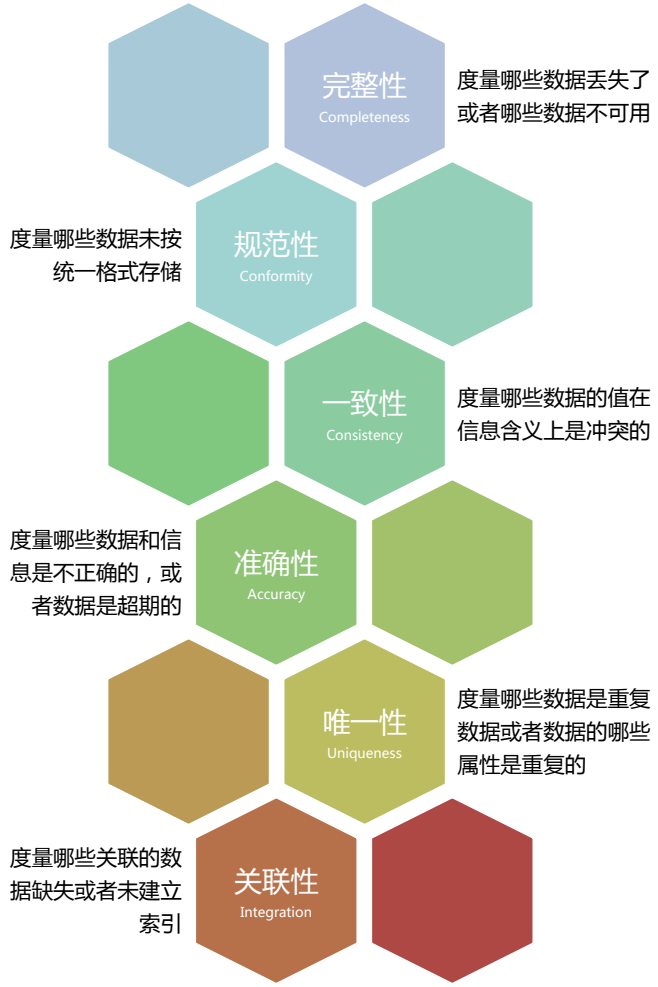
大数据质量提升的系统性方案

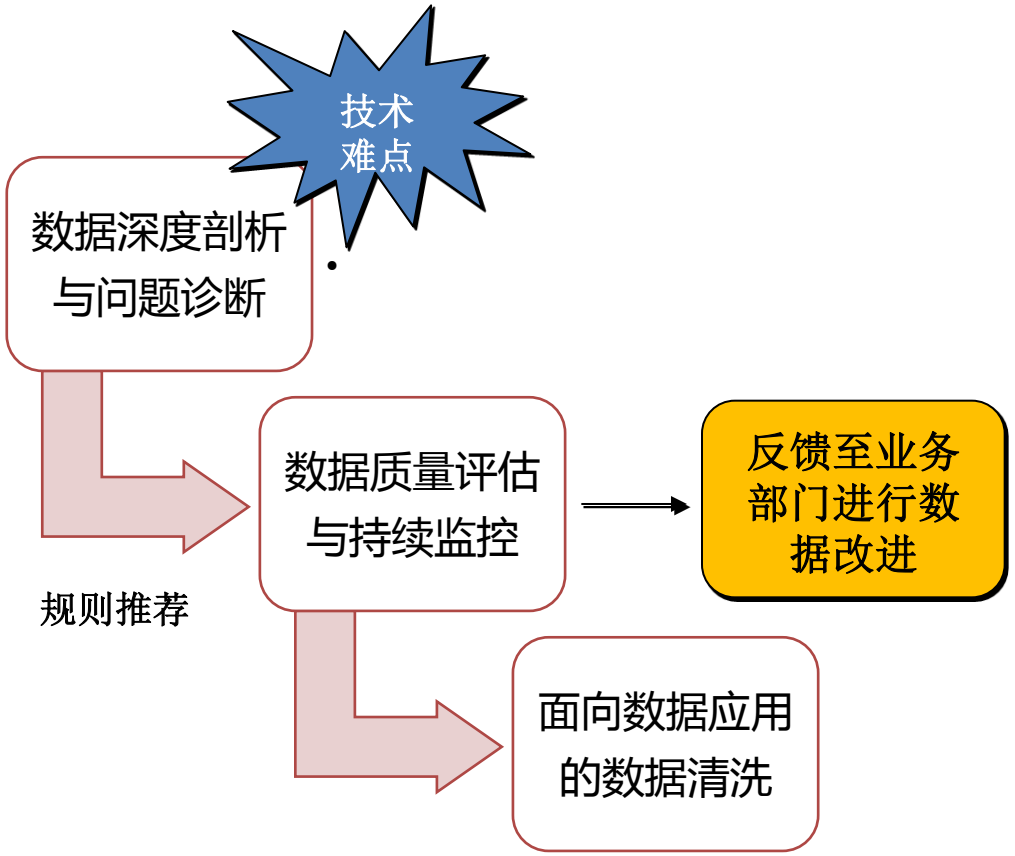
大数据清洗治理系统

国家电网
STATE GRID
全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

大数据质量提升的系统性方案

数据质量（Data Quality）是数据分析结论有效性和准确性的基础，也是最重要的前提和保障。“数据质量”尚无统一定义，但学术界普遍认可以下评价维度：





国家电网
STATE GRID

全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

大数据质量提升的系统性方案

数据质量管理建议 → 质量提升方案 → 数据剖析 → 数据清洗

理解 | 分析 | 清洗 | 监控 | 维持

一次性 | 连续性

物资
人资
OMS
CMS
PMS
95598
自动化
用采

数据剖析、检核评估及清洗

业务改进

数据剖析、检核评估及清洗

结构分析和映射

探查并分析多个源业务系统 → 将数据映射到目标结构

各源系统专责

持续清洗和监控数据质量

跨系统专责

审计和报告主数据质量

建立企业数据质量规则

部署规则交互/批量/实时

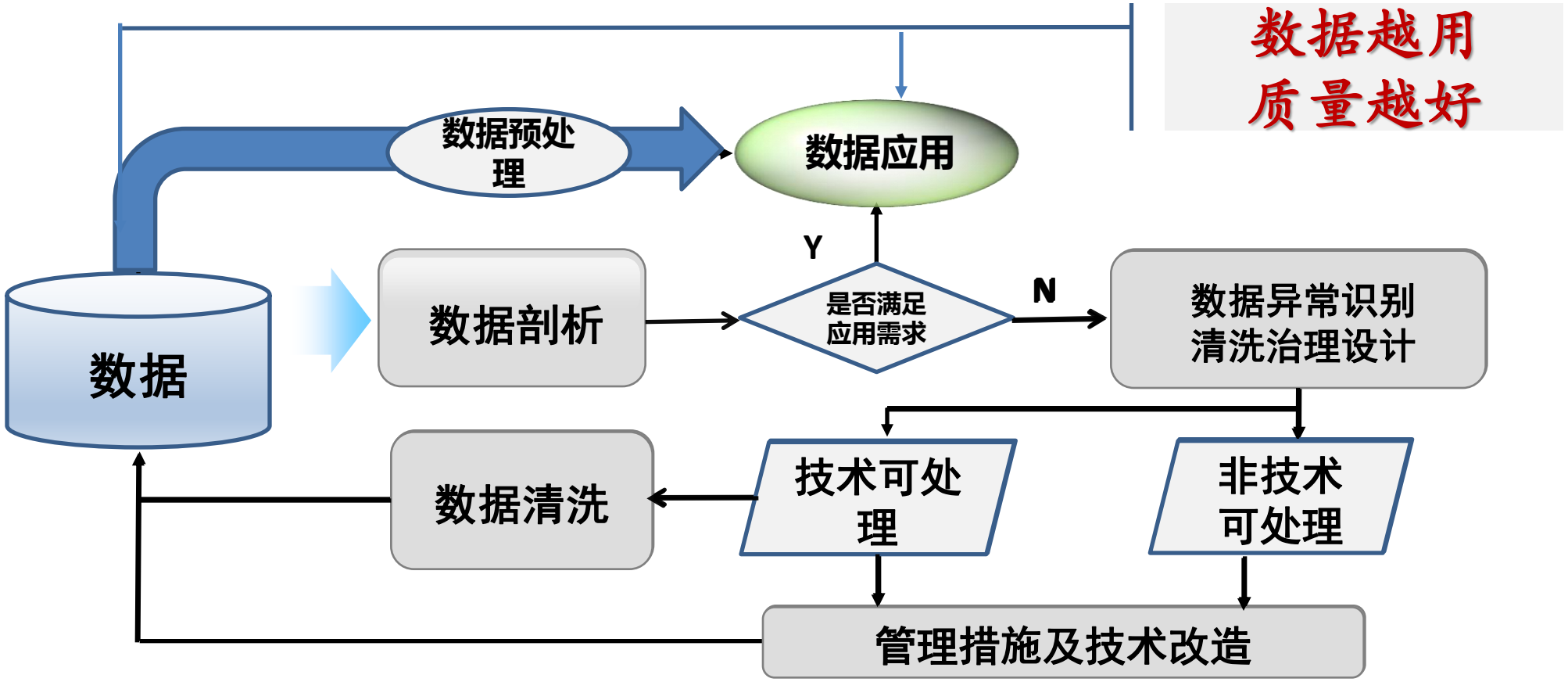
将问题报告发送至源系统

规范和制度体系

国家电网
STATE GRID
全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

大数据质量提升的系统性方案

数据质量管理建议 → 质量提升方案 → 数据剖析 → 数据清洗



国家电网
STATE GRID
全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

大数据质量提升的系统性方案

数据质量管理建议 → 质量提升方案 → 数据剖析 → 数据清洗

主数据模型设计优化

业务逻辑及架构顶层设计优化

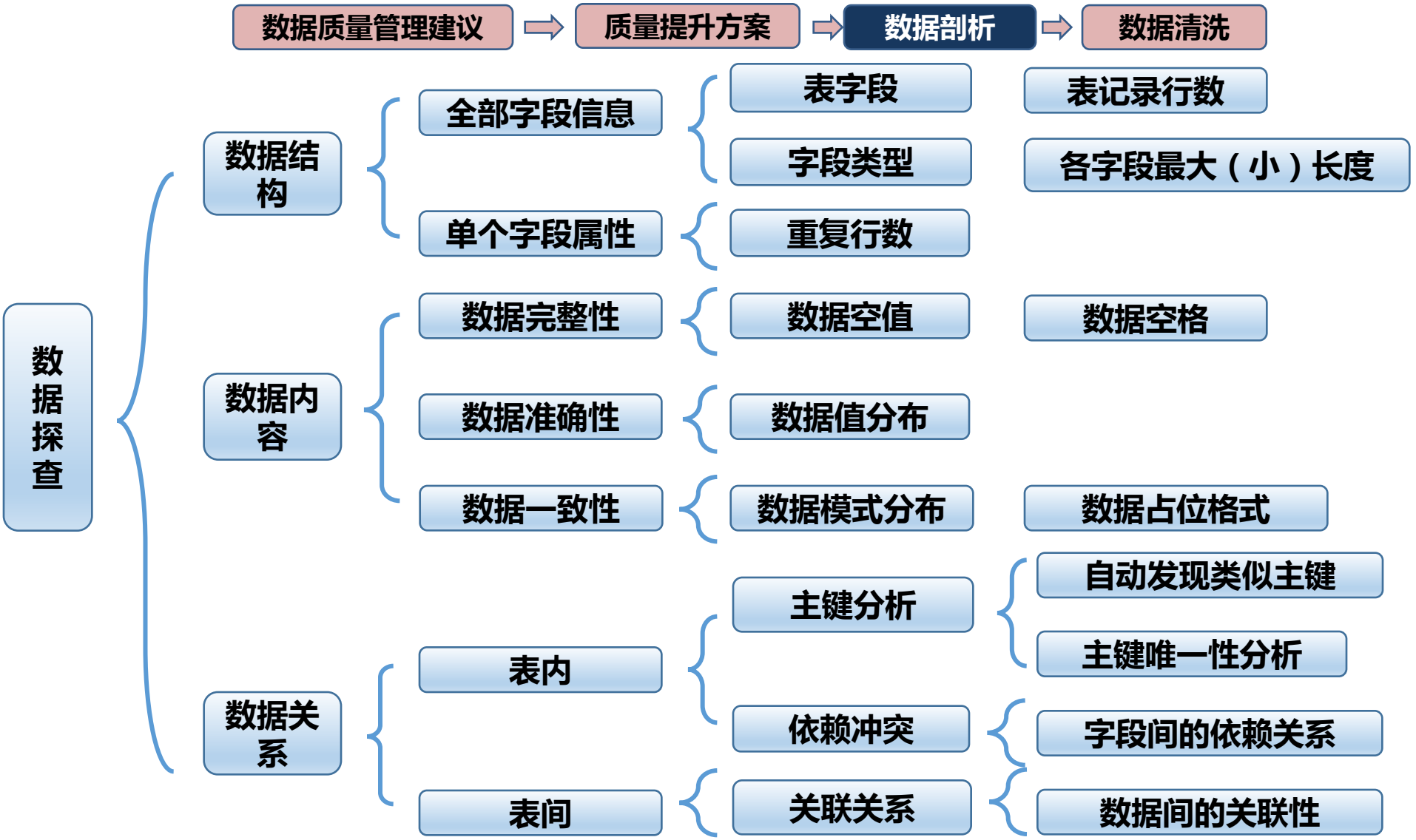
业务数据库结构优化

促进

数据应用及共享

国家电网
STATE GRID
全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

大数据质量提升的系统性方案



国家电网
STATE GRID
全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

大数据质量提升的系统性方案

数据质量管理建议 → 质量提升方案 → 数据剖析 → 数据清洗

数据

探查

评估

问题清单

- ✓ 采集或监测类连续数值型数据缺失
- ✓ 用户或设备属性缺失
- ✓ 类型信息错误或不规范
- ✓ 数据量纲错误
- ✓ 设备或用户地址不规范
- ✓ 数据值异常或不一致
- ✓ ……

数据质量提升目标及规则

算法模型清洗
- 插值法
- 预测
- 回归
- 聚类
- 分类
- 文本分析

替换和计算
- 值替换
- 表内计算
- 跨表计算

国家电网
STATE GRID

全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

目 录

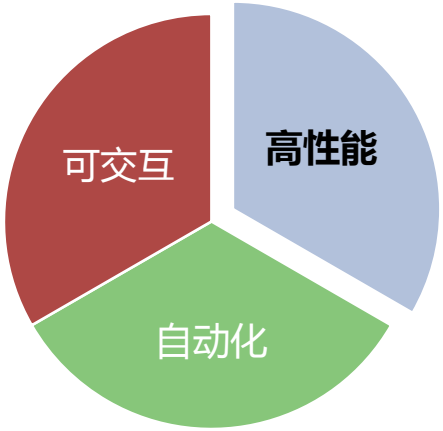
全球能源互联网与大数据

大数据质量提升的系统性方案

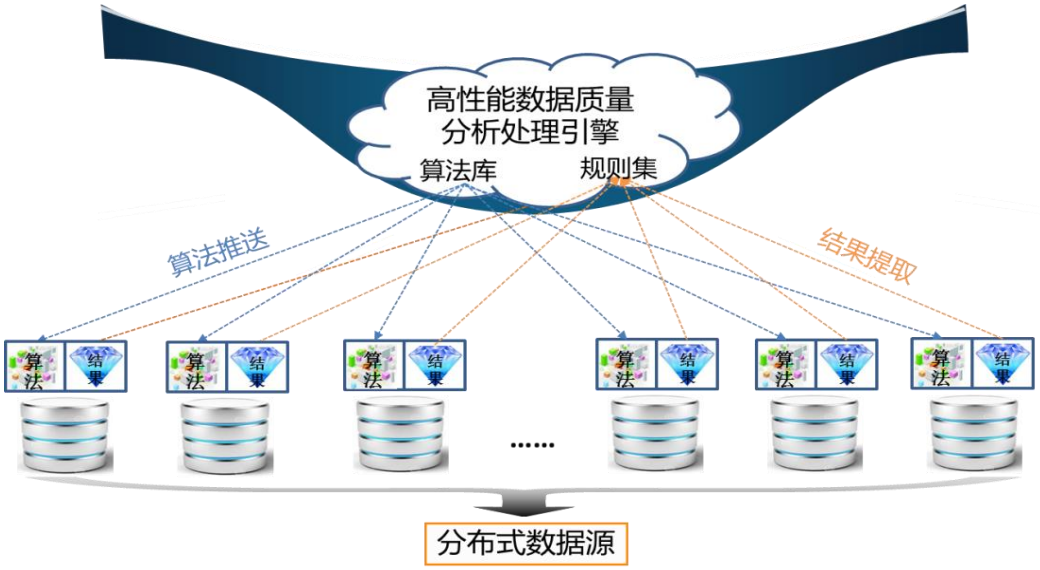
大数据清洗治理系统

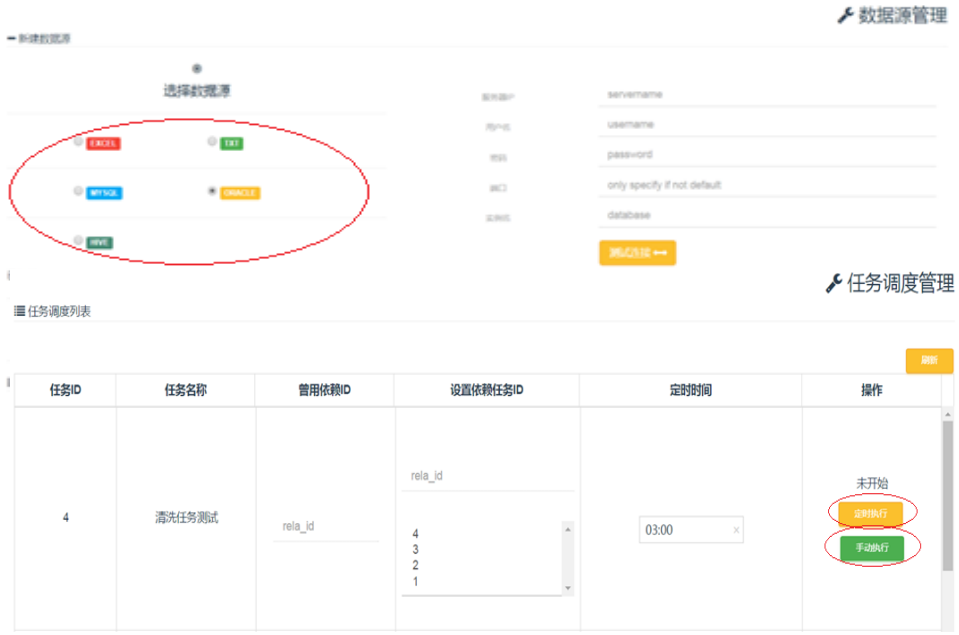
国家电网
STATE GRID
全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

大数据清洗治理系统-功能特性



- 多源异构数据源的集成
- 灵活配置规则，内嵌基于业务逻辑的清洗模型并支持扩展
- 基于大数据的批量数据清洗，可持续调度执行
- 灵活部署，高效执行





国家电网
STATE GRID
全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

大数据清洗治理系统-核心功能

基于规则的清洗治理

解决的问题

1、替换标准内容
2、主表与码表关联获取码值内容
3、统一字段格式及内容
4、空值字段赋值处理
5、特殊字符替换
6、多列间逻辑运算或拼接
7、删除重复行记录

规则

表达式

解决的问题

1、提取有规律的信息（如提取数字）
2、逻辑运算或简单统计信息
3、数据精度标准化
4、日期标准化、结构化（如分解年月日及季度等）
5、字符串大小写，去空格、特殊字符
6、正则替换（如替换数字开头的内容）
7、特殊格式数据解析（如URL、json格式数据的解析）
8、补足位数，简单加密字段内容

基于模型的清洗治理

解决的问题

1、地址标准化
2、地址补全、地址去重
3、错误行政区划所属关系纠正、行政区划错字纠正
4、门牌楼号正则化
5、基于地址丰富数据维度（增加市、区、镇）

文本型算法模型

数值型算法模型

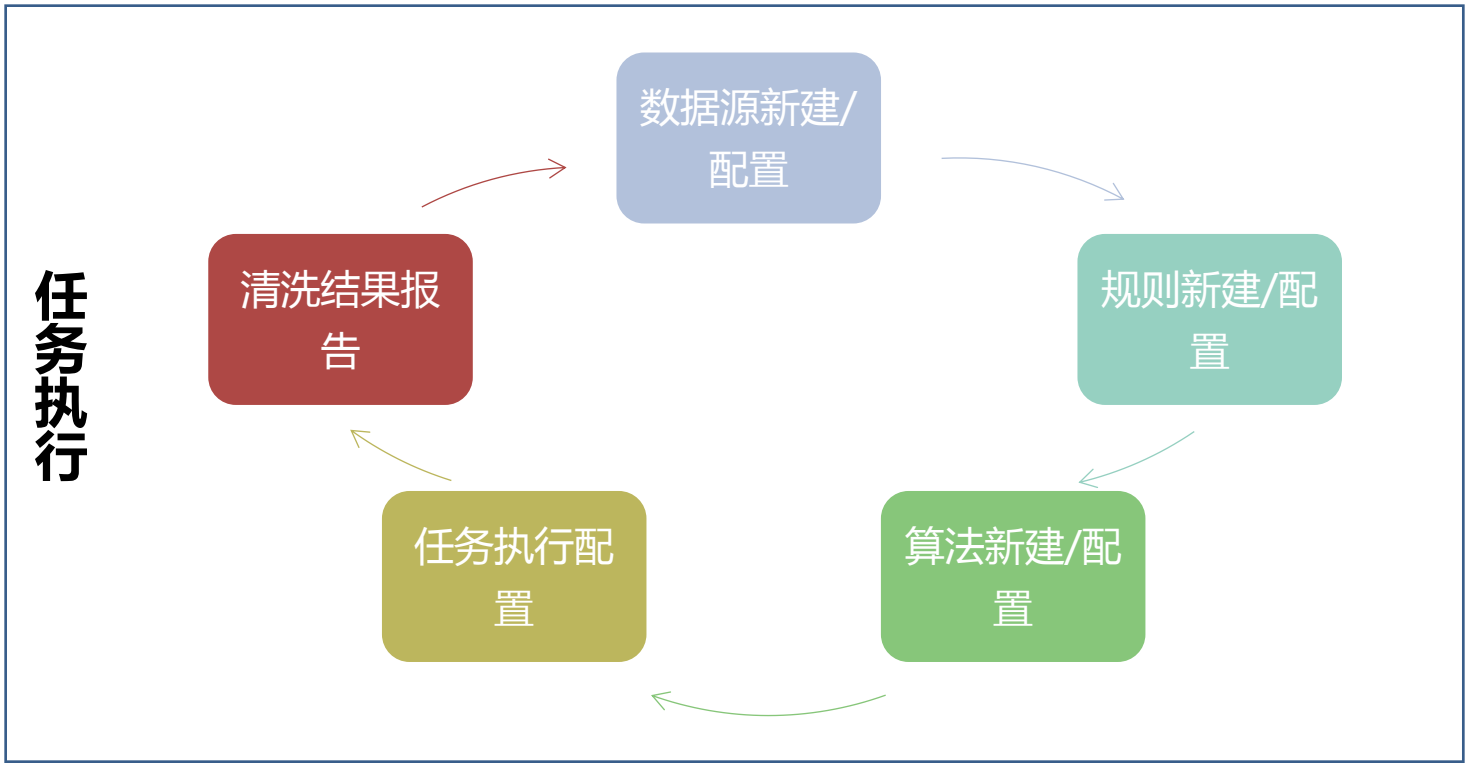
解决的问题

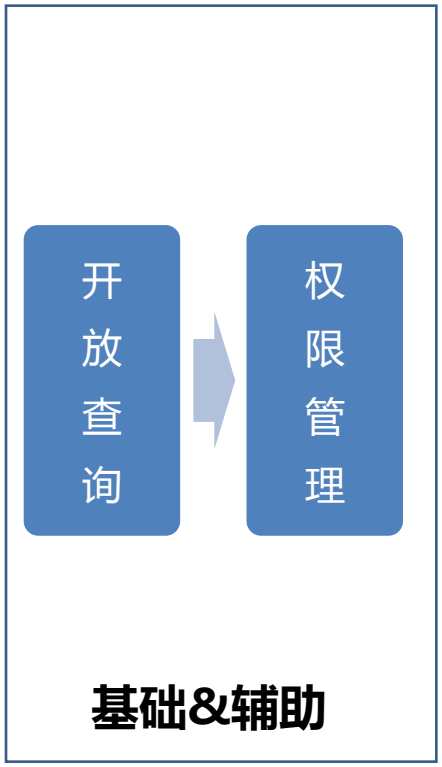
1、最优匹配搜索CBMS：适用于数据矩阵的缺失值填补,对连续缺失有优势，能够清洗平均连续丢失10个数值型数据
2、改进的交替式最小二乘优化IALS：适用于时间序列类数据的缺失值的预测填补
3、时间序列ARIMA模型与卡尔曼滤波KARM：适用于时间序列数据的缺失值填补,对分散性缺失比较有优势，能够填补分散性缺失50%的数值型数据
4、集合模型：将多个独立缺失值还原算法的结果相结合，返回更准确的估算结果

国家电网
STATE GRID
全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

大数据清洗治理系统-应用流程

通过简单的配置操作即可完成专业而持续的数据清洗治理，使经过简单训练的人员轻松完成繁杂的数据处理。





国家电网
STATE GRID

全球能源互联网研究院
GLOBAL ENERGY INTERCONNECTION RESEARCH INSTITUTE

谢谢！