长江众创
WWW.CJZC.NET.CN

打造数据中央厨房 助力大数据创业

上海长江时代众创空间数字技术有限公司
孙繁荣

引言

- 大数据为新财富，价值堪比石油

——世界经济论坛报告

- 数据列入企业资产负债表只是时间问题

——维克托

- 一个组织如果没有认识到管理数据和信息如同管理有形资产一样极其重要，那么他在新经济时代将无法生存

——汤姆.彼得斯

数据创业维艰：从零到一

创业公司往往面临着招人难、融资难、推广难等等各种困境。

在瞬息万变的移动互联网领域，创业公司要想在巨头的夹缝中求生存，

高效的产品研发能力和快速更新迭代，才是生存与发展的关键。

聚焦产品创新

据统计：一个数据分析项目中，数据获取，数据清洗等准备工作占据了将近70%的时间。

数据准备包括：数据获取、数据清洗、存储归档等

其他工作包括：数据观察、数据建模、数据挖掘等

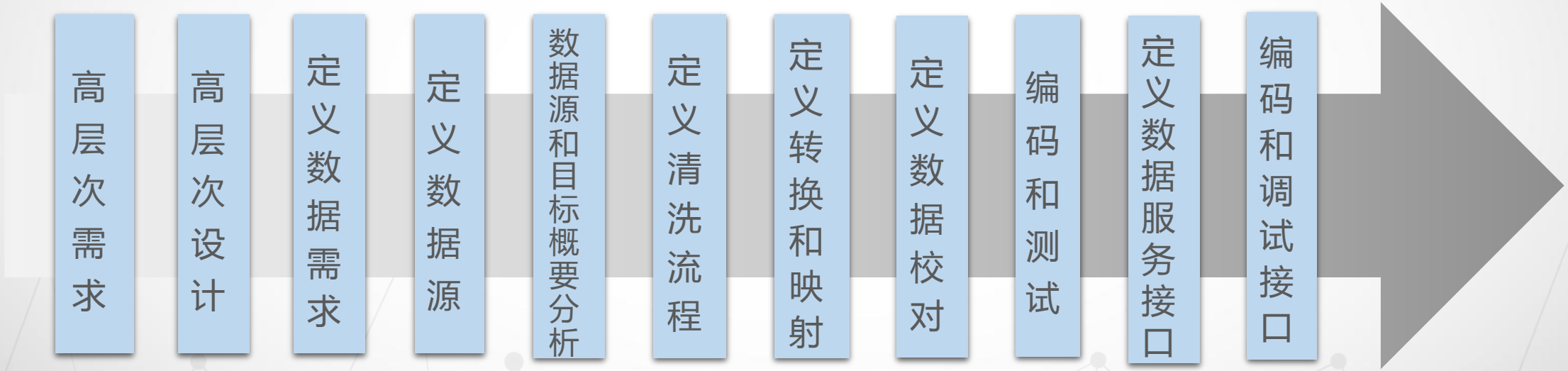
企业可以聚焦在专业领域，把数据获取和清洗环节作为产品的外延服务

数据AI生态圈

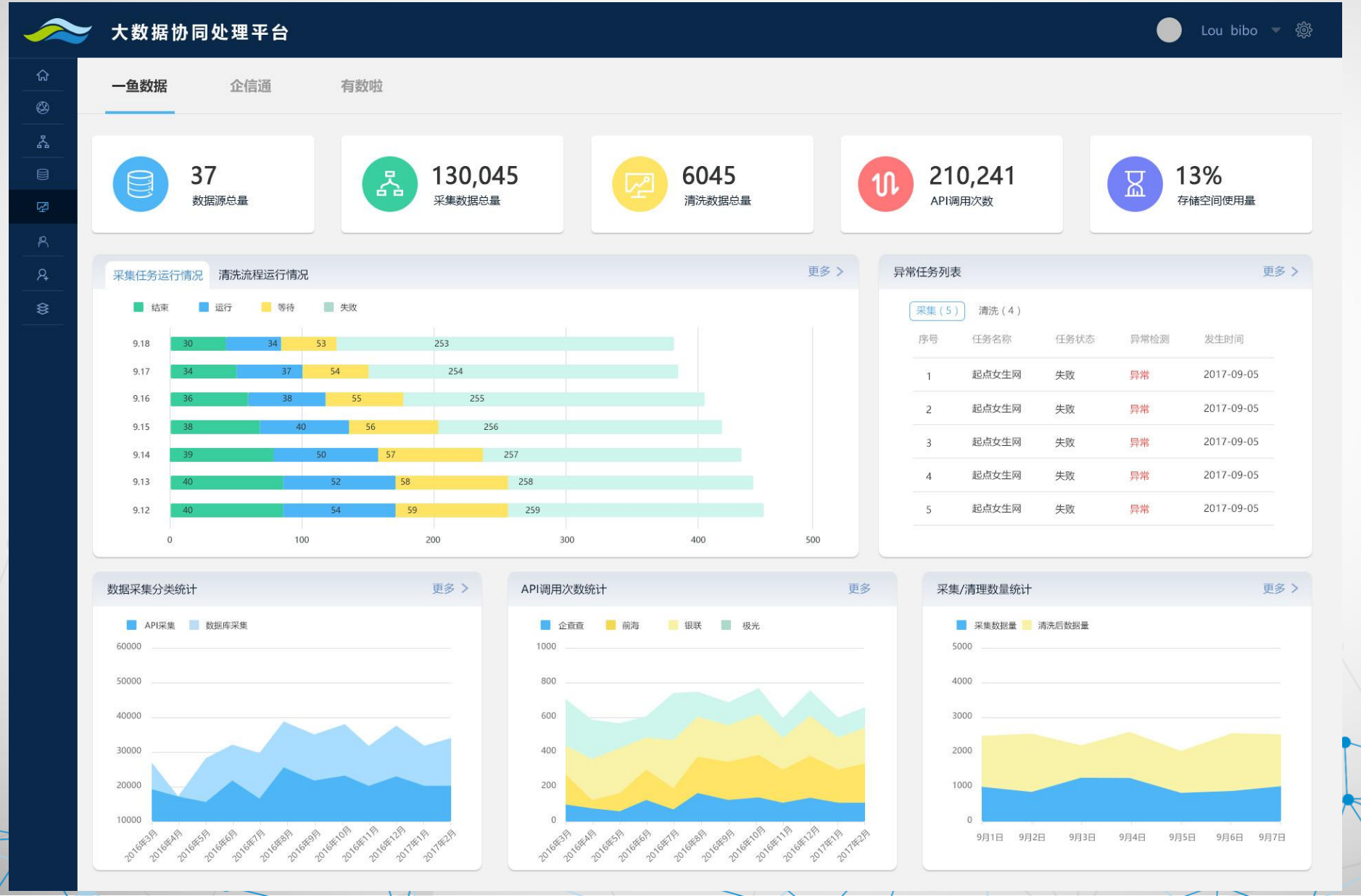
长江众创打造数据协同工作平台，提供“数据中央厨房”服务，整合多方资源，孵化各类垂直场景金融数据及人工智能企业建设AI生态圈

数据集成开发流程

数据集成开发遵循完整的软件项目开发流程



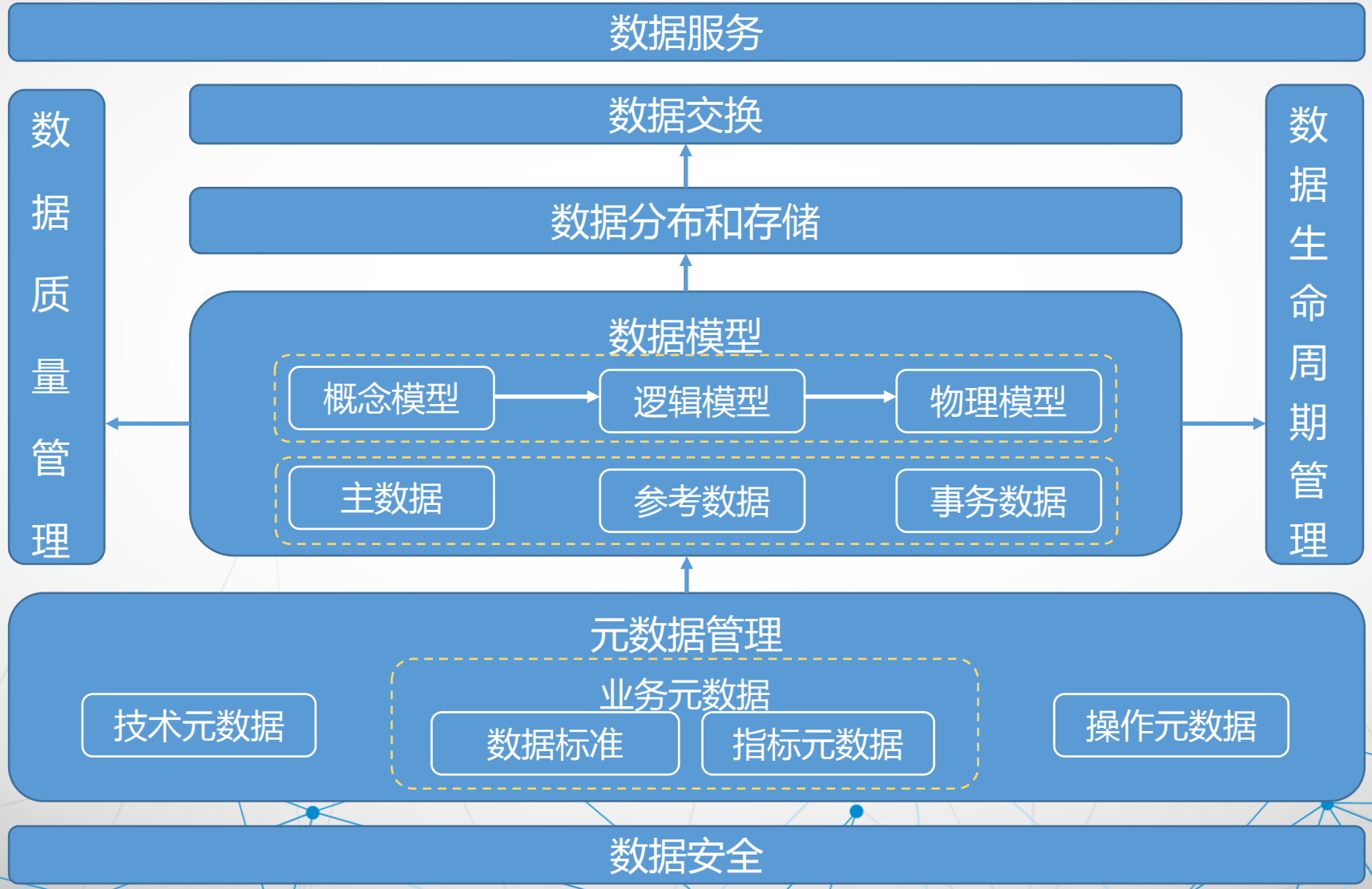
数据协同工作平台一览


大数据协同处理平台
Lou bibo
一鱼数据
企信通
有数啦
37
数据源总量
130,045
采集数据总量
6045
清洗数据总量
210,241
API调用次数
13%
存储空间使用量
采集任务运行情况
清洗流程运行情况
更多 >
结束
运行
等待
失败
9.18 30 34 53 253
9.17 34 37 54 254
9.16 36 38 55 255
9.15 38 40 56 256
9.14 39 50 57 257
9.13 40 52 58 258
9.12 40 54 59 259
0 100 200 300 400 500
异常任务列表
更多 >
采集 (5)
清洗 (4)
序号 任务名称 任务状态 异常检测 发生时间
1 起点女生网 失败 异常 2017-09-05
2 起点女生网 失败 异常 2017-09-05
3 起点女生网 失败 异常 2017-09-05
4 起点女生网 失败 异常 2017-09-05
5 起点女生网 失败 异常 2017-09-05
数据采集分类统计
更多 >
API采集
数据库采集
60000
50000
40000
30000
20000
10000
2016年3月 2016年4月 2016年5月 2016年6月 2016年7月 2016年8月 2016年9月 2016年10月 2016年11月 2016年12月 2017年1月 2017年2月
API调用次数统计
更多
企查查
前海
银联
极光
1000
800
600
400
200
0
2016年3月 2016年4月 2016年5月 2016年6月 2016年7月 2016年8月 2016年9月 2016年10月 2016年11月 2016年12月 2017年1月 2017年2月
采集/清理数量统计
更多 >
采集数据量
清洗后数据量
5000
4000
3000
2000
1000
0
9月1日 9月2日 9月3日 9月4日 9月5日 9月6日 9月7日

数据集成面临的挑战

1. 多源异构数据源的接口复杂性
    1. 多样的应用数据库
    2. 不同外部数据供应商的接入协议（通信、数据格式、认证、加密、字典）
    3. 多种类型的数据类型：结构化、半结构化、非结构化
2. 语义歧义
    1. 同一概念在不同数据源的涵义不同
    2. 不同概念在不同数据源的涵义相似
3. 实例歧义
    1. 数据记录唯一标示
    2. 关联数据识别
4. 数据标准和质量不规范
    1. 格式、编码、度量单位、缺值、多值等
    2. 指标、统计口径不一致
5. 多源数据的更新频率/方式不一致

数据管理架构应对复杂的数据集成



数据协同工作平台四大特点

- 支持多源异构数据源快速接入
- 元数据配置驱动的采集和智能清洗流程
- 基于容器的微服务架构支持高并发及海量数据存储
- 数据全生命周期管理与溯源

数据集成框架

数据源

数据采集

- 数据源管理
- 采集任务调度管理
- 采集监控
- RawData

数据预处理

- 清洗流程管理
- 智能数据清洗
- 数据质量控制
- FactData

数据集市

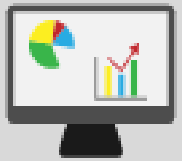
- 数据模型管理
- 数据服务接口管理
- 数据同步管理
- AppData

- 元数据管理
- 权限管理
- 用户管理
- 数据管理

运营管理

- 运行监控
- 资源调度
- 日志管理
- 多租户管理

行业应用



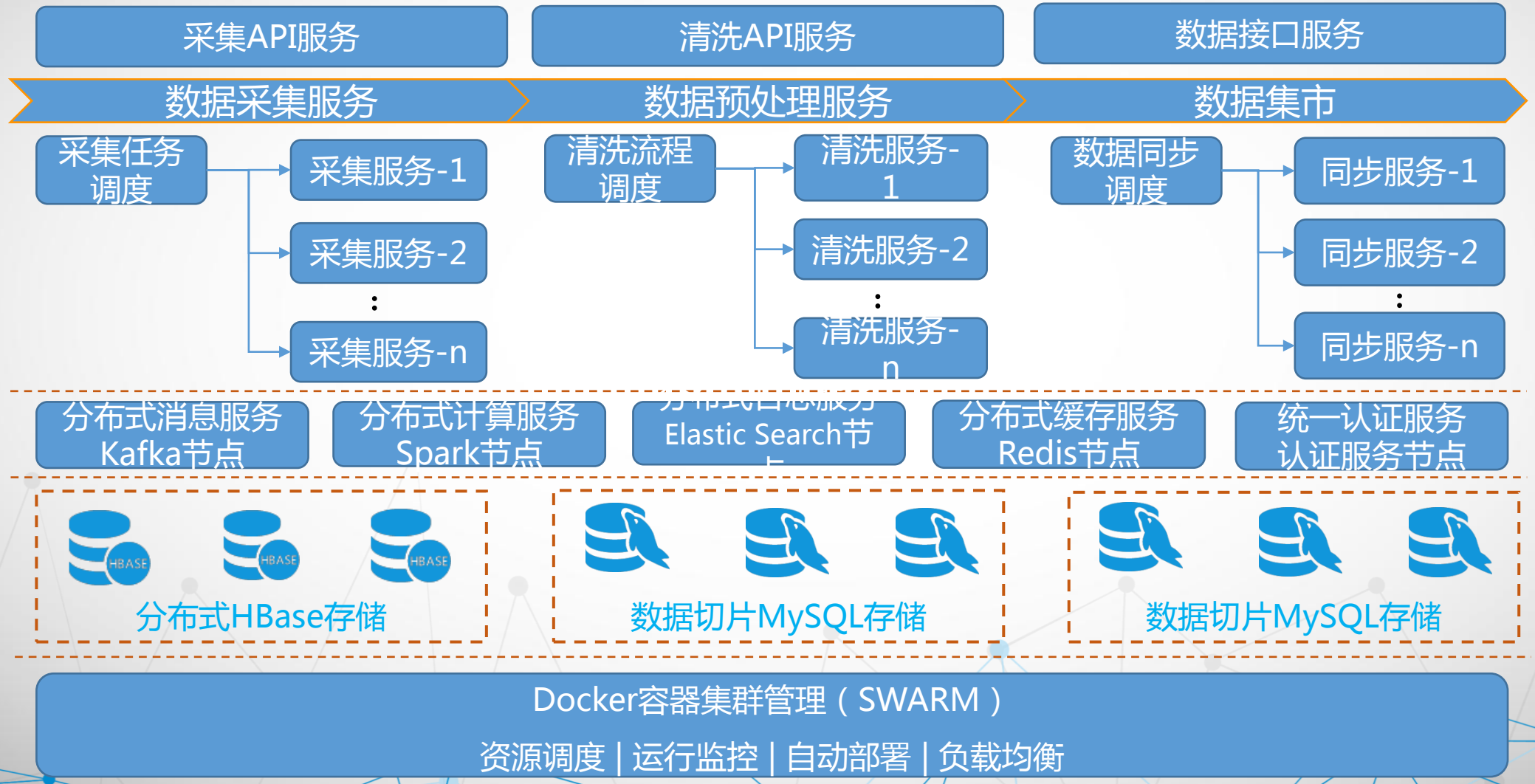
一鱼数据 YIYUDATA

金时征信 GOLDEN ERA

数据协同工作平台关键技术指标

存储量	• 千亿量级数据 • PB级空间容量
吞吐量	• 日均亿级数据处理 • 秒级单次采集及清洗
可靠性	• 所有数据保留三份全量副本 • 动态资源调度应对高峰压力

技术架构



关于我们

About us

打造数据中央厨房，助力数据企业孵化

秉承 MOM 理 念，共建AI产业生态圈

产品与服务

公司形成以数据服务为核心

以咨询顾问、社群服务、投资孵化服务为支撑

相互协同的业务格局

咨询顾问

- 为传统行业中的企业提供大数据+相关咨询顾问

数据服务

- 场景金融数据服务
- 数据“中央厨房”

创投孵化

- 关联数据公司及AI职能的孵化和投资

社群服务

- 数据爱好者社群服务
- 数据咨询师社群服务
- 企业家社群服务

数创空间站

长江众创积极响应国家“大众创业，万众创新”的号召

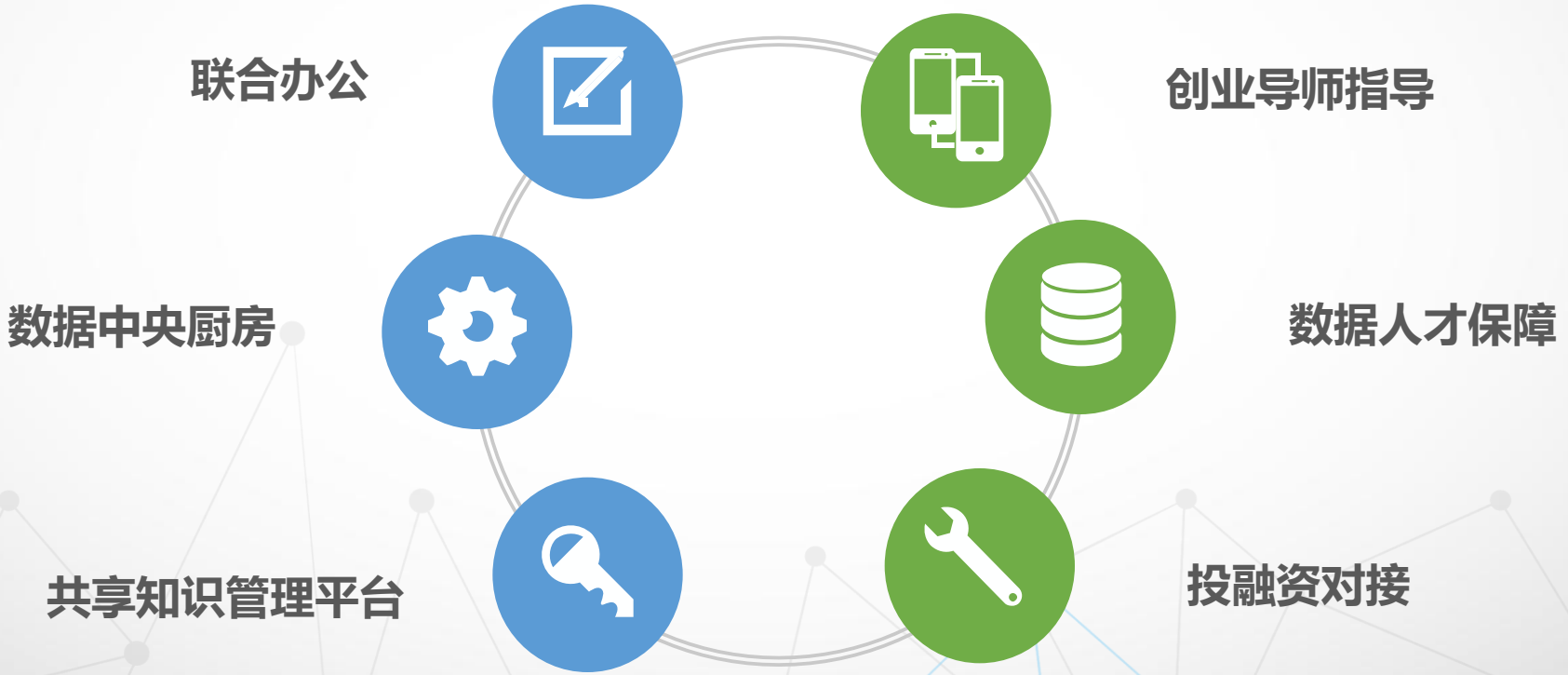
在上海、北京、宁波等地建立数创客空间（Dataker Station）,助力大数据&人工智能产业创新升级

其中上海的空间站为浦东新区创新型孵化器

数创空间站

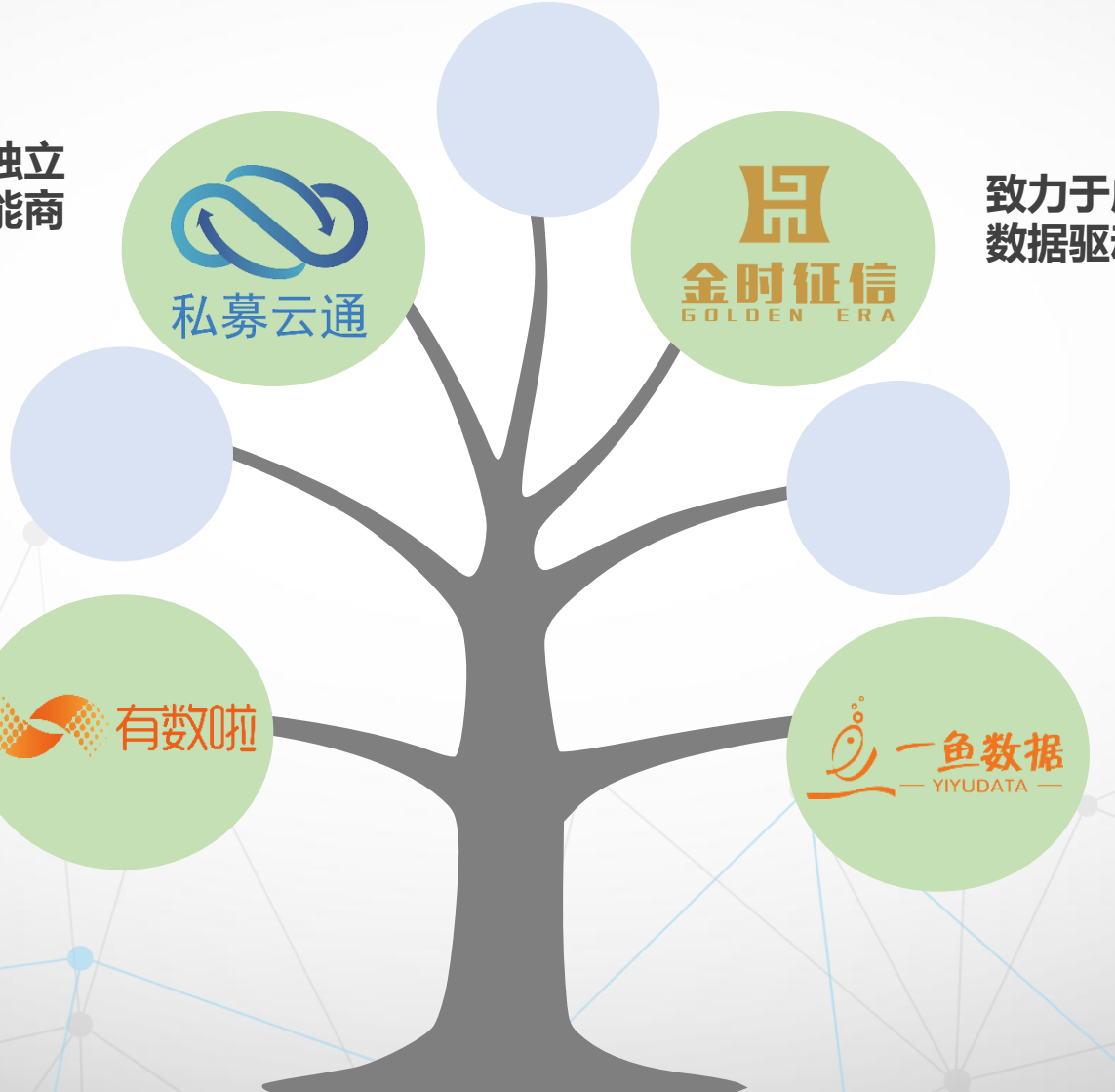
为大数据及人工智能创新企业提供办公空间与人才聚集平台,大数据技术支持的多样化全流程服务体系

同时为各阶段创业者提供投融资对接服务，释放入孵企业发展潜能,加速入孵企业成长,帮助创业者走好每一步



数创空间站

目前，数创空间站已经孵化了公司内外数个大数据创新项目，助力大数据产业创新升级。



致力于成为中国一流的独立基金信息服务和配置赋能商

致力于成为公信力卓越的数据驱动型征信服务平台

致力于成为最IN的数据客在线社群

致力于成为中国领先的泛娱乐大数据服务商

长江光启创新学院

学院以数据“三师”培训体系为核心，整合大数据、AI各界专家，通过线上大数据Live及线下沙龙等形式为入孵企业及大数据爱好者提供人才培养与实训，同时为有志创业人才提供创业进阶指导。

大数据职业技能培训

大数据 分析师课程

- 数据可视化
- 大数据统计分析基础
- 贝叶斯分析与应用
- 数据库与数据处理
- 网络爬虫与文本挖掘
- 数据治理
- 深度学习
- R语言编程与开发
- 大数据平台技术与应用
- Python机器学习

大数据 工程师课程

- Java核心开发
- Hadoop应用
- 大数据行业应用导论
- 大数据架构
- 微服务与实现
- Hopping精讲
- 大数据仓库Hive精讲
- Spark与机器学习
- 分布式存储:HDFS、Hbase

大数据 咨询师课程

- 大数据导论
- 产业研究方法
- 大数据实践与应用
- 数据治理
- 项目管理与有效沟通

- 大数据统计分析实战训练营(R/Python)
- Web全栈工程师数据可视化训练营
- 量化金融分析师实战训练营

大数据Live

数创客

THANK YOU