ArchData

技术峰会成都站

主办方：中生代技术 FRESHMAN TECHNOLOGY 快CTO 互联网创业技术服务平台

2018年1月27日成都菁蓉国际广场3W COFFICE 蓉漂茶馆

关于我自己：

拥有20+年软件产品研发经验，在商业智能/基于大数据的机器学习，分布式服务集群系统架构设计等软件产品的项目组织，开发，管理方面拥有丰富经验。文斌在2016年9月加入客如云，担任技术副总裁，负责客如云产品研发，技术架构及餐饮大数据平台的研发和管理工作。在加入客如云之前，文斌担任OpenText Analytics /Actuate VP of Engineering，负责 Information Hub/Big Data Analytics等下一代基于大数据的商业智能平台的研发工作。在Actuate，文斌还领导了Actuate的商业智能（Business Intelligence）旗舰产品Actuate iServer的研发工作。文斌还是开源商业智能项目Eclipse BIRT（Business Intelligence Reporting Tool）的主要设计者。

基于大数据的商业智能/数据挖掘的架构演进及客如云的大数据实践

何文斌
客如云技术VP

目录

商业智能（Business Intelligence）简史

- 1865，Richard Miller Devens 使用business intelligence 这个词来描述银行家 Sir Henry Furnese成功的原因： He had an understanding of political issues, instabilities, and the market before his competitors。
- 1958，IBM ‘s Hans Peter Luhn （商业智能之父） 发表文章 A business intelligence system，定义为： an automatic system developed to disseminate information to the various sections of any industrial, scientific, or government organization。
- 1989， Gartner analyst Howard Dresner，提出用business intelligence 来概括 decision support systems (DSS）和executive information system (EIS)
- 1990s， business intelligence 1.0 = ETL + DSS + EIS
- 在21世纪的头十年，以数据仓库为平台，报表/仪表盘/数据分析为核心服务的Business Intelligence 2.0
- 在最近10年，以Hadoop/DakeLake为平台，数据挖掘/机器学习，自服务为核心服务的商业智能（有人也叫BI 3.0）

Business Intelligence 2.0

BI 2.0 Landscape

BI 提供商

数据库提供商

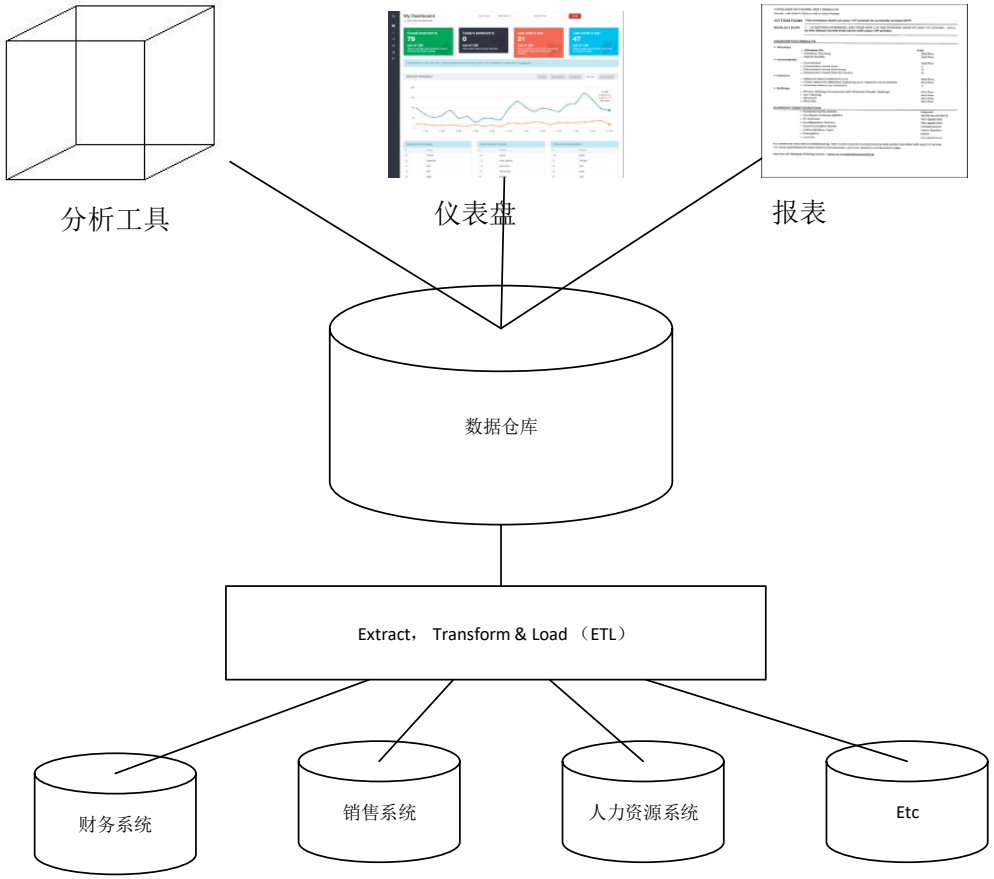
ETL 提供商

BI 2.0 - 关系型数据仓库架构

A central repository of integrated data from one or more data sources。

Reporting & Analysis

Data Governance



BI 2.0 应用场景评估

应用场景	支持
公司报表/Corporate reporting	✓
公司报表打印/Pixel Perfect Reporting	✓
即席数据分析/Ad-hoc analysis	✓
实时数据分析/Real-Time Analytics	☒
数据挖掘/预测/Advanced Analytics	☒
全数据分析/All Data Analytics	☒
自服务商业智能/Self-service BI	☒

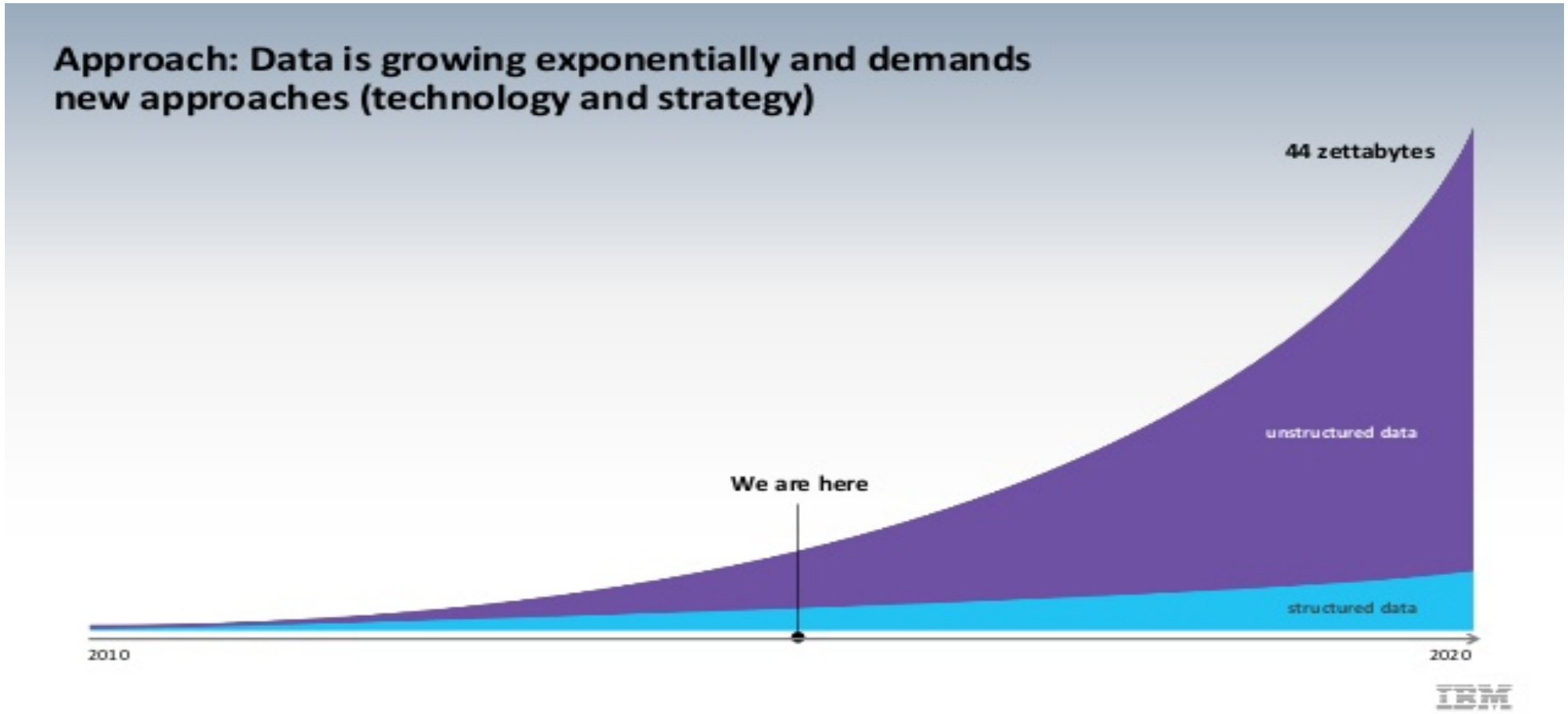
数据类型：

结构化数据

特性	评价
敏捷	low
扩展性	low
成本	high
性能	Low -》 middle
一致性	high
适应性	low
安全	High

大数据挑战

数据容量的爆炸性增长



* Source from Internet

BigData Questions?

How to implement recommendation or anomaly detection achieving **Low Latency**?

Consumer Intelligence agent

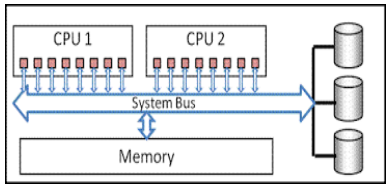
How to enable data science/advanced analytics team for **predictive and advanced analytics**?

Data Scientist/ Analyst

How to provide real-time dashboards or self-service BI with **high quality and good performance o**ver terabytes and petabytes of data?

Business users

Extended BI 2.0 Landscape



Massively Parallel Processing (MPP)

Cloud platform

TERADATA

hp VERTICA

IBM NETEZZA

SQL Server 2012 Parallel Data Warehouse

Google BigQuery

snowflake

OpenGeo Suite | Pivotal Greenplum Database

SAP HANA

Amazon Redshift

Extended BI 2.0架构

数据源

- 数据库 结构化数据
- Text/XML 半结构化数据
- text/image /video 非结构化数据

数据集成

- ETL
- Messaging
- API
- Replication

数据存储

- 数据仓库
- 数据集市
- Operational data stores

数据分析

- 查询&报表
- OLAP Cubes
- 高级数据分析

数据展示

- 浏览器
- 移动设备
- Web Services
- PC

Extended BI 2.0应用场景评估

应用场景	支持
公司报表/Corporate reporting	✓
公司报表打印/Pixel Perfect Reporting	✓
即席数据分析/Ad-hoc analysis	✓
实时数据分析/Real-Time Analytics	✓ ✗
数据挖掘/预测/Advanced Analytics	☒
全数据分析/All Data Analytics	☒
自服务商业智能/Self-service BI	☒

数据类型：

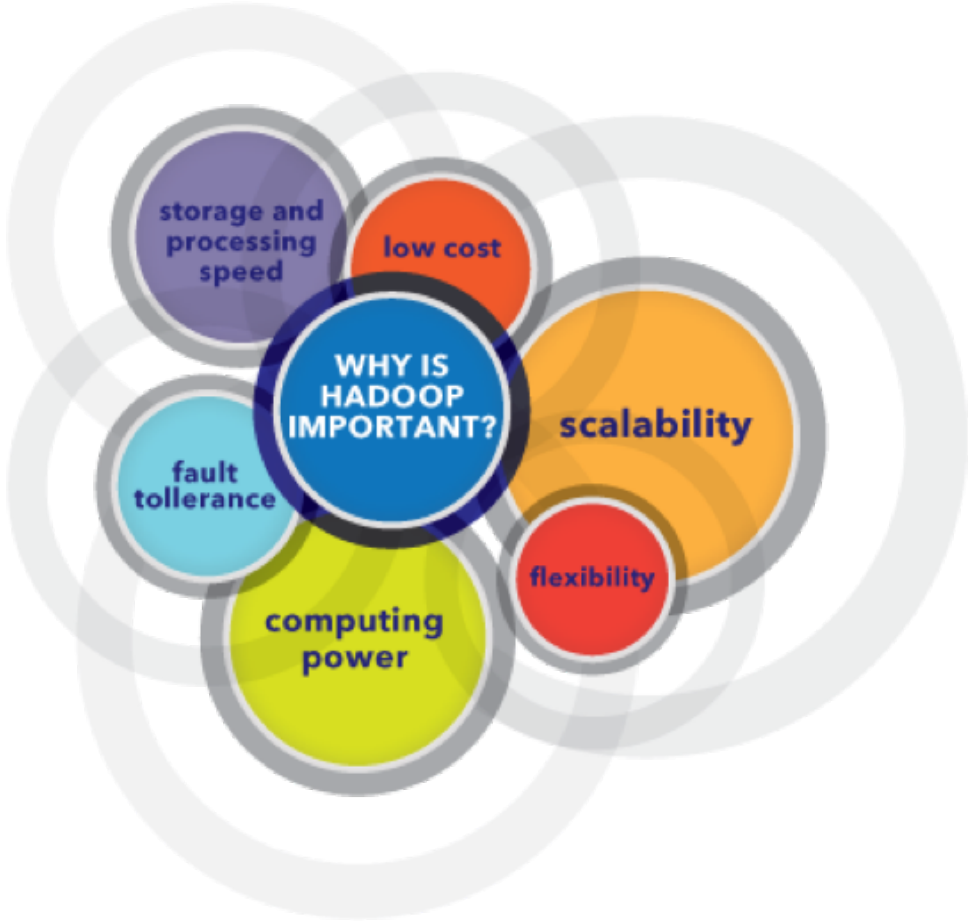
结构化数据

半结构化数据

非结构化数据

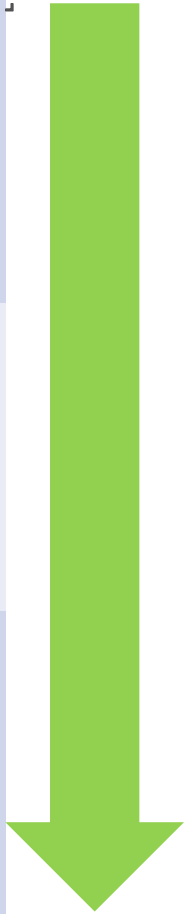
特性	评价
敏捷	low
扩展性	middle
成本	Very high
性能	middle
一致性	middle
适应性	middle
安全	High

我们需要一个敏捷，低成本， 可扩展的解决方案来应对互联网应用所带来的大数据挑战！

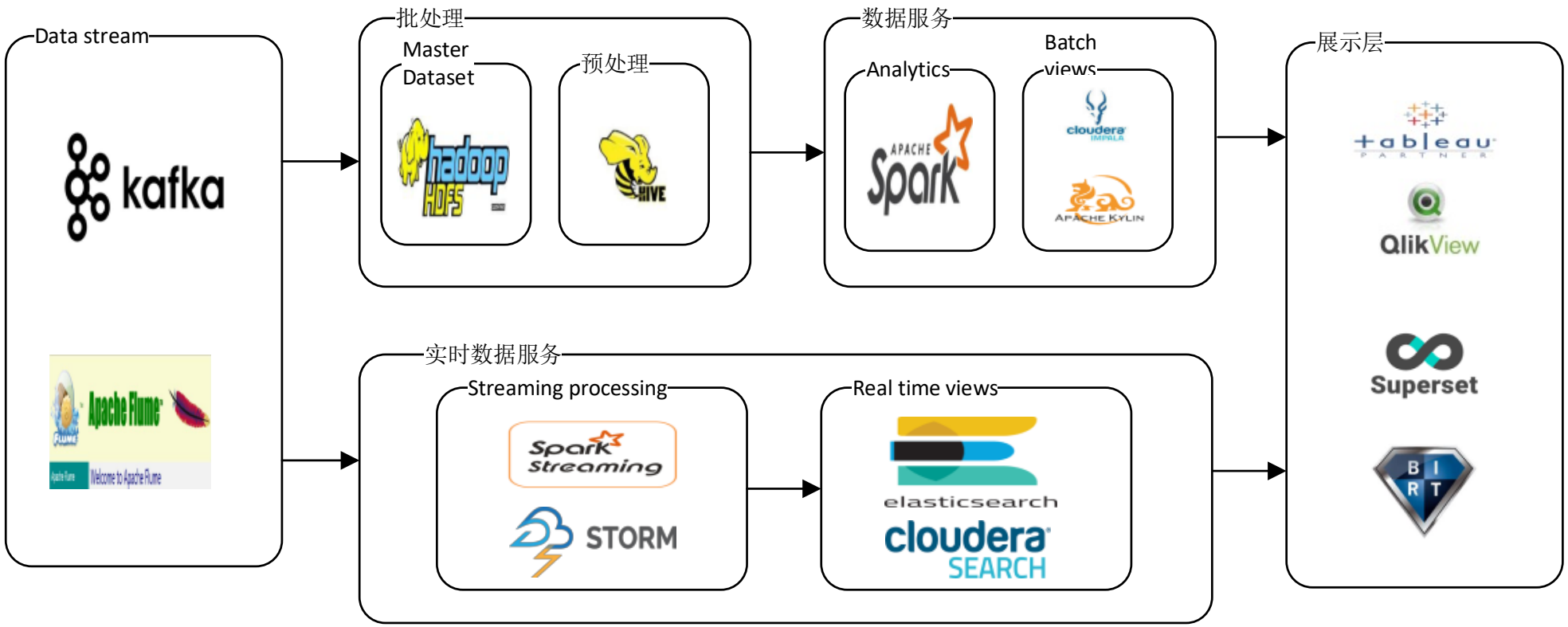


Bigdata 商业智能的架构演进

	数据源	提取方法	存储方法	探索方法	消费
传统数据仓库	ERP database High utility data Curated datasets	MQ Series Informatica	EDW OLAP cubes	Custom solution Parametrized reports	MSTR, Cpgnos, Actuate/BIRT,etc Static reports Rigid Dashboards
V1-Hadoop MR	Logs/Semi structured data Machine generated data	Open Sources ETL：Kettle,Sqoop Flume, Custom MR	HIVE/Hbase/HDFS MPP Search	BI on Hadoop	Custom application
V2-Spark on HDFS	Events & Sensor data Stream generated data	Kafka Storm Spark Streaming	Drill, Hive on Tez Impala, SparkSQL Casandra, Kudu	ZoomData Python/R/Scala	Custom application Integrated analytics



Lambda 流式计算架构



什么是数据湖？与数据仓库的不同

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.

Data Lake应用场景评估

应用场景	支持
公司报表/Corporate reporting	☒
公司报表打印/Pixel Perfect Reporting	☒
即席数据分析/Ad-hoc analysis	✓
文本分析/text Mining	✓
数据挖掘/预测/Advanced Analytics	✓
全数据分析/All Data Analytics	✓
自服务商业智能/Self-service BI	✓

数据类型：

结构化数据

半结构化数据

非结构化数据

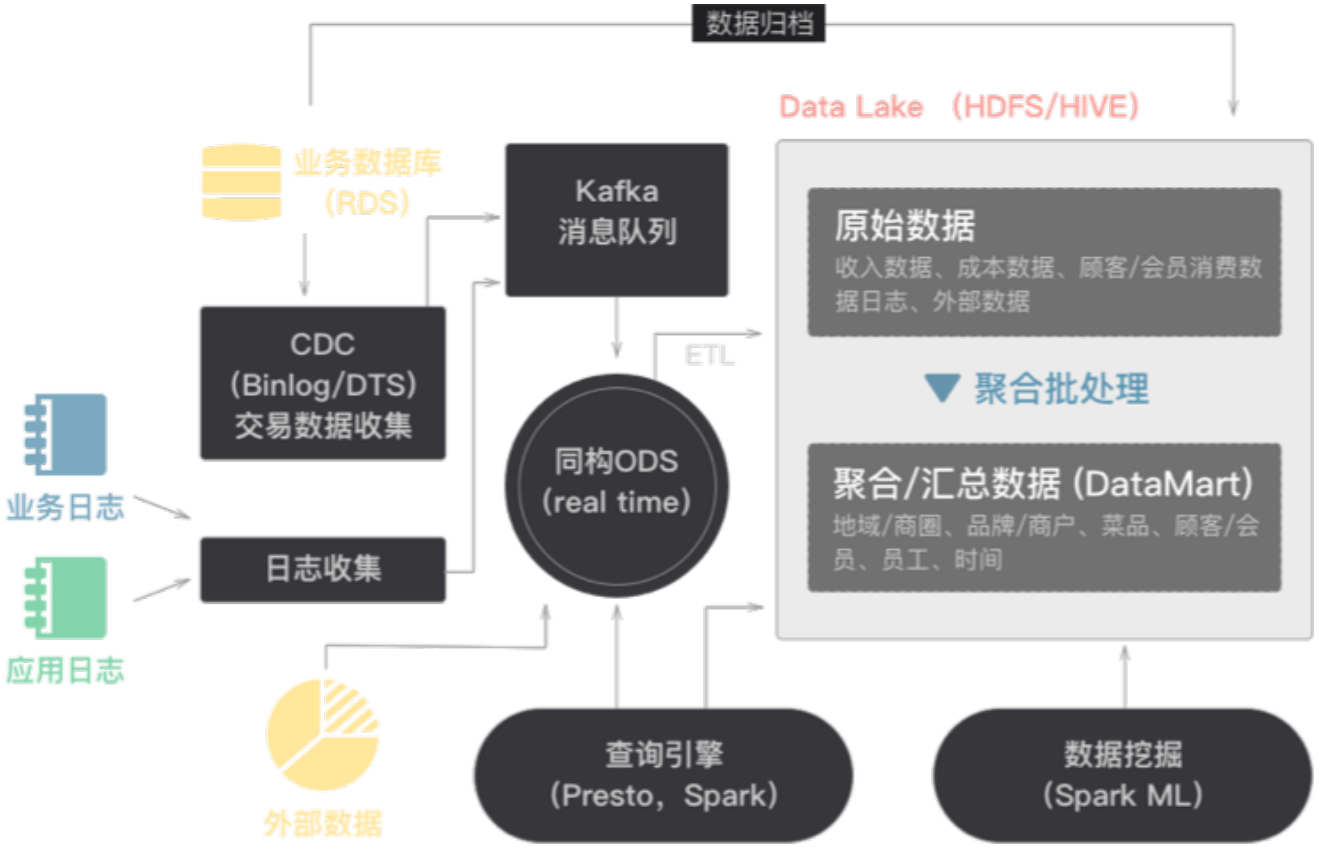
特性	评价
敏捷	High
扩展性	High
成本	middle
性能	high
一致性	low
适应性	high
安全	middle

但真实的情况是，我们两者都需要：

数据仓库（processed data）来生成各类报表/仪表盘，支持公司的各个业务部门

数据湖（raw data），供数据科学家来做数据挖掘（data mining/discovery）

客如云大数据平台架构



强大的数据采集、处理、存储、挖掘、分析功能，自动生成包括餐厅、顾客帐户信息，记录餐饮消费数据，顾客对餐厅、菜品、服务的评价，以及服务员对顾客的CRM描述，等等

完善的数据分析模型，自动生成商户经营数据模型，顾客行为/消费模型，智能连接商户与第三方服务比如金融，供应链采购，以及外卖服务等

基于机器学习的数据挖掘/智能推荐引擎，帮助商户经营者更好地做出经营决策，更合理调整门店布局，菜品升级/改良，精准营销。为消费者提供推荐服务

客如云大数据挖掘实践

基于文本分析的菜品标准化分类

	需要标准化的菜品	标准菜	相似度
2	湖南红烧肉	红烧肉	0.97851
3	稻草红烧肉	红烧肉	0.96228
4	三品坊红烧肉	红烧肉	0.94559
5	毛氏红烧肉。	红烧肉	0.96844
6	南昌红烧肉	红烧肉	0.97257
7	乡里红烧肉	红烧肉	0.97802
8	5来个红烧肉sz	红烧肉	0.95997
9	红烧肉（3块起售）	红烧肉	1
10	陈年花雕红烧肉（位）	红烧肉	0.95068
11	席面红烧肉	红烧肉	0.94753
12	瓦罐红烧肉	红烧肉	0.97475
13	外婆烧的红烧肉	红烧肉	0.86567
14	东北红烧肉	红烧肉	0.98353
15	红烧肉炒饭	红烧肉炒饭	1
16	粉条红烧肉	红烧肉炖粉条	0.94144
17	红烧肉土豆盖饭.	红烧肉盖饭	0.90687
18	红烧肉土豆盖饭	红烧肉盖饭	0.90687
19	红烧肉豆腐盖饭	红烧肉盖饭	0.92869
20	尖椒红烧肉盖饭	红烧肉盖饭	0.87036
21	红烧肉卤面	红烧肉卤面	1
22	蜀爷·红烧肉面	红烧肉面	0.98708
23	红烧肉粉/面	红烧肉面	0.92421
24	红烧肉丝豆腐	红烧肉末豆腐	0.97443
25	红烧肉烧豆角双拼.	红烧肉烧豆角	0.9945
26	招牌红烧肉圆	红烧肉圆	1
27	红烧肉圆（单点商品不含米饭）	红烧肉圆	1

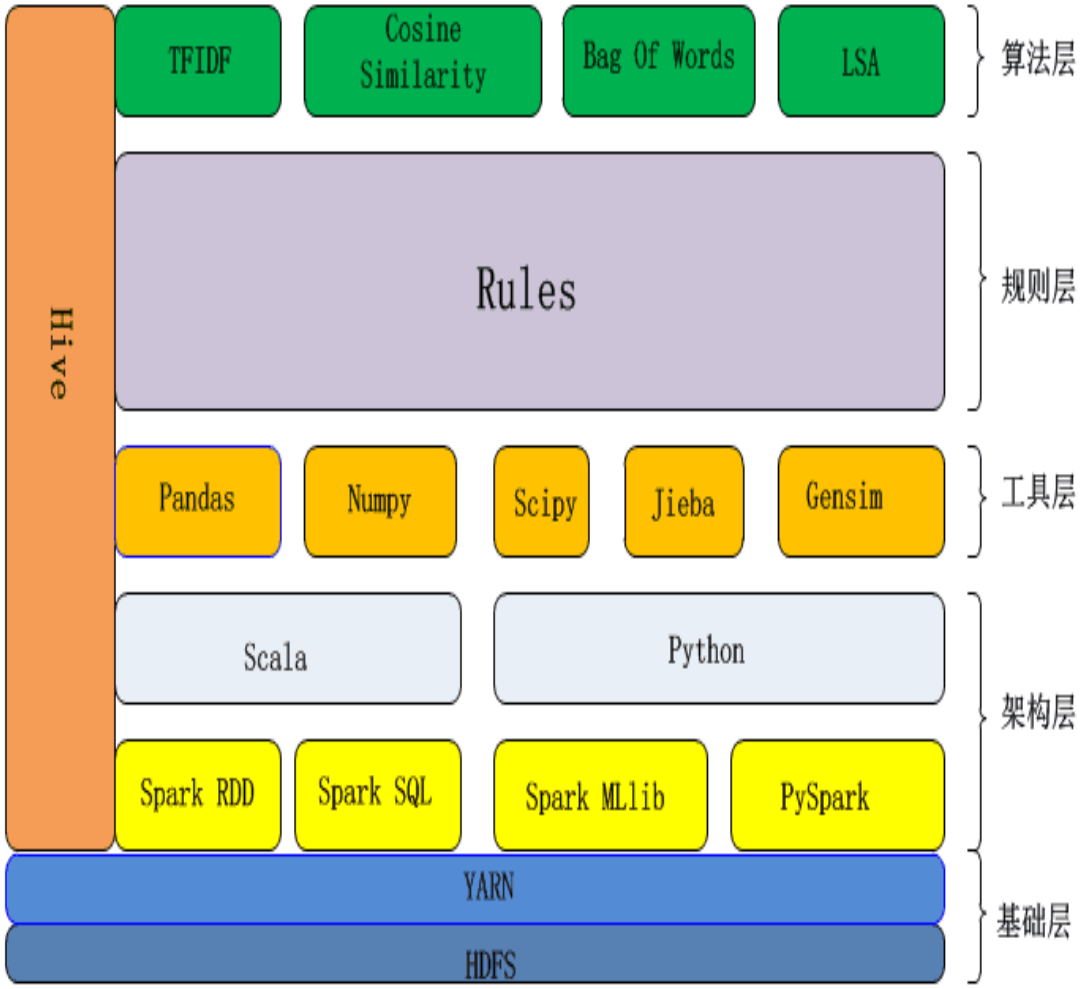
注：相对于模糊匹配，不会将红烧肉盖饭，红烧肉面，红烧肉圆等匹配为红烧肉

基于文本分析的菜品标准化分类-1

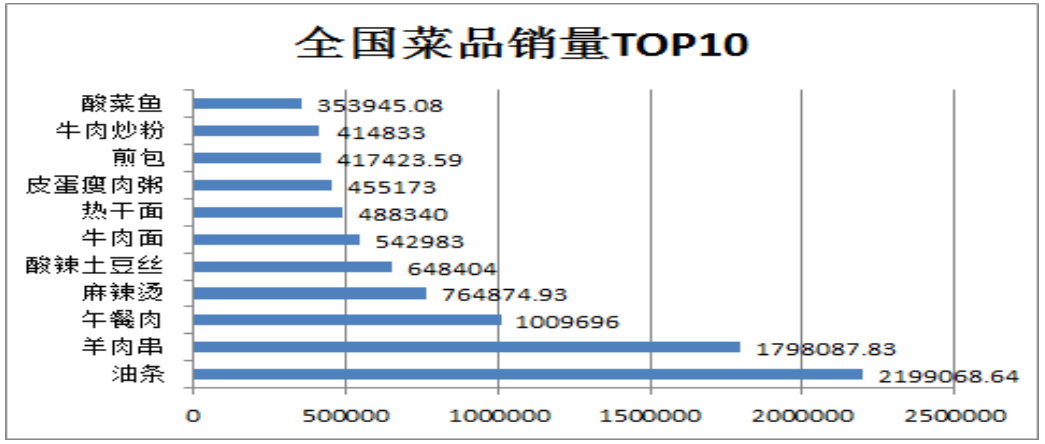
	A	B	C
1	需要标准的菜	标准菜	相似度
2	白菜猪肉水饺（12个）	猪肉白菜水饺	1
3	白菜猪肉水饺（赠送）	猪肉白菜水饺	1
4	猪肉白菜水饺1两6个	猪肉白菜水饺	0.98941
5	小份猪肉白菜水饺	猪肉白菜水饺	1
6	切丁白菜猪肉手工水饺	猪肉白菜水饺	0.87889
7	白菜猪肉水饺20个+卤蛋+可乐	猪肉白菜水饺	1
8	水饺　猪肉白菜	猪肉白菜水饺	1
9	白菜猪肉水饺大	猪肉白菜水饺	1
10	猪肉白菜水饺	猪肉白菜水饺	1
11	白菜猪肉水饺	猪肉白菜水饺	1
12	白菜猪肉水饺(中)	猪肉白菜水饺	1
13	白菜猪肉水饺（18个）	猪肉白菜水饺	1
14	湾仔码头白菜猪肉水饺720g	猪肉白菜水饺	0.9604
15	猪肉白菜水饺20个	猪肉白菜水饺	1
16	猪肉白菜水饺大	猪肉白菜水饺	1
17	（外）白菜猪肉水饺	猪肉白菜水饺	1
18	干捞猪肉白菜水饺	猪肉白菜水饺	0.98433
19	猪肉白菜水饺/20个	猪肉白菜水饺	1
20	白菜猪肉水饺/20个	猪肉白菜水饺	1
21	白菜猪肉水饺(大)	猪肉白菜水饺	1
22	白菜猪肉水饺特价	猪肉白菜水饺	1
23	切丁白菜猪肉水饺	猪肉白菜水饺	0.97073
24	白菜猪肉水饺【小份】	猪肉白菜水饺	0.9605
25	手工猪肉白菜水饺（20个）	猪肉白菜水饺	0.90043
26	中份猪肉白菜水饺	猪肉白菜水饺	0.98772
27	猪肉白菜水饺（大份）	猪肉白菜水饺	1

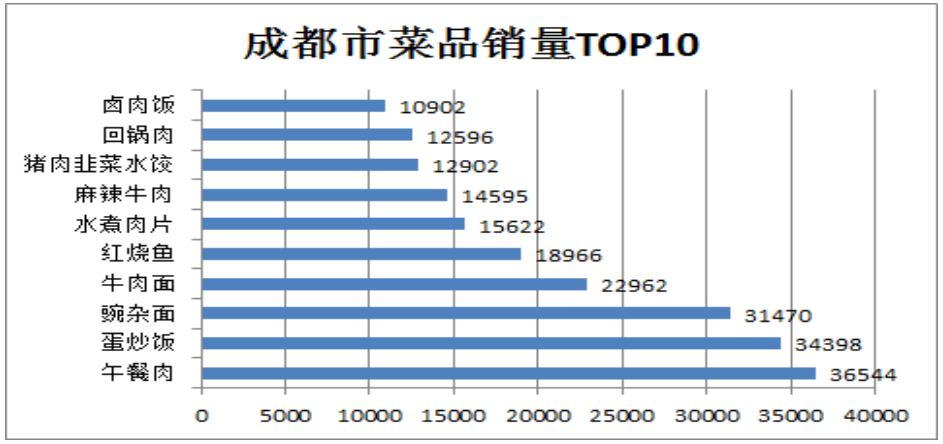
注：各种猪肉白菜和白菜猪肉类水饺均可以匹配出来

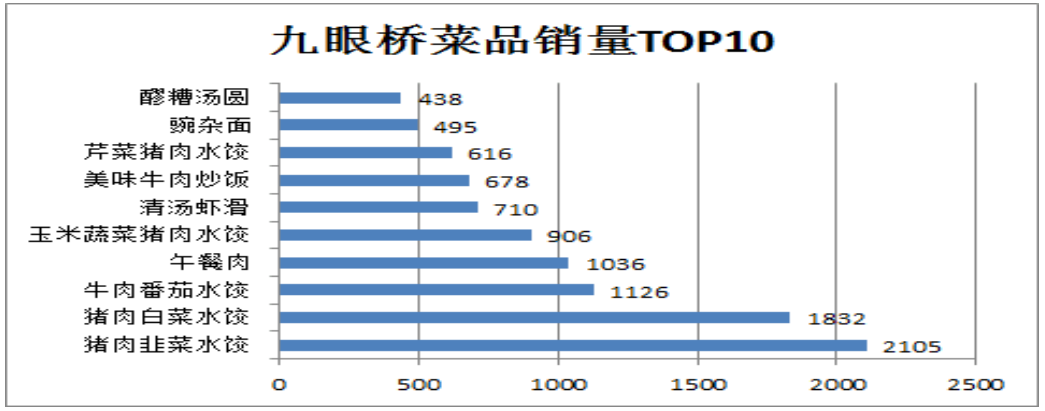
菜品标准化分类的应用架构



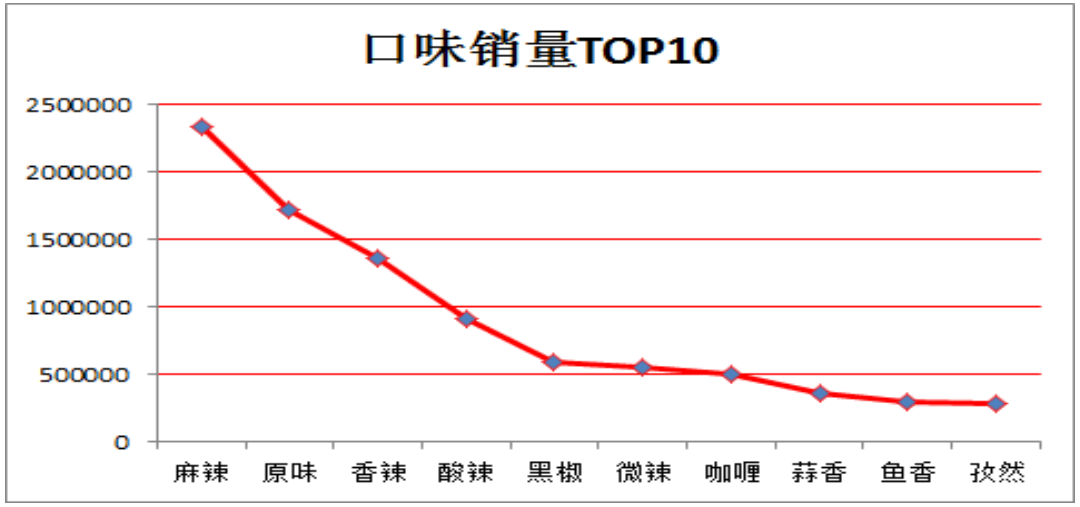
基于标准化菜品分类的菜品销量统计

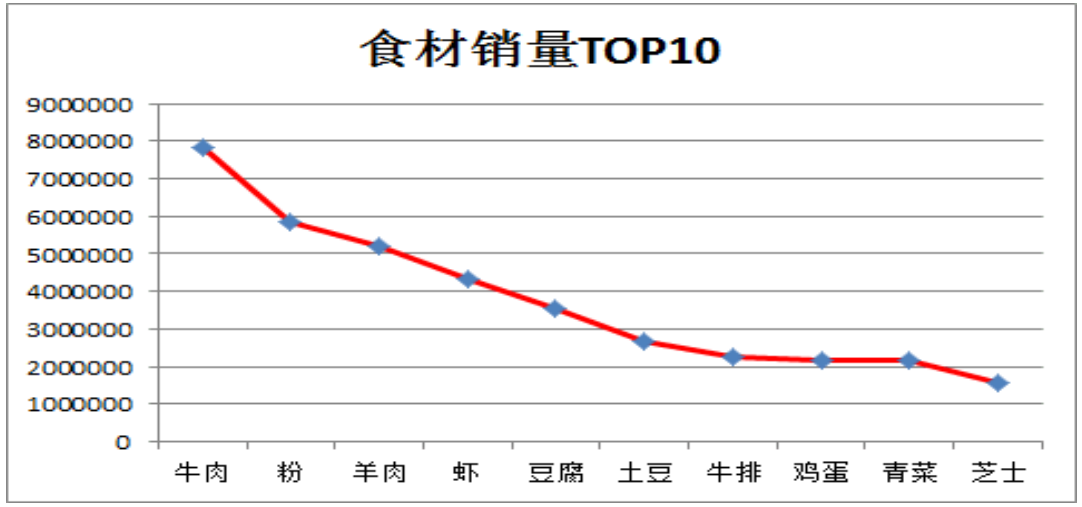




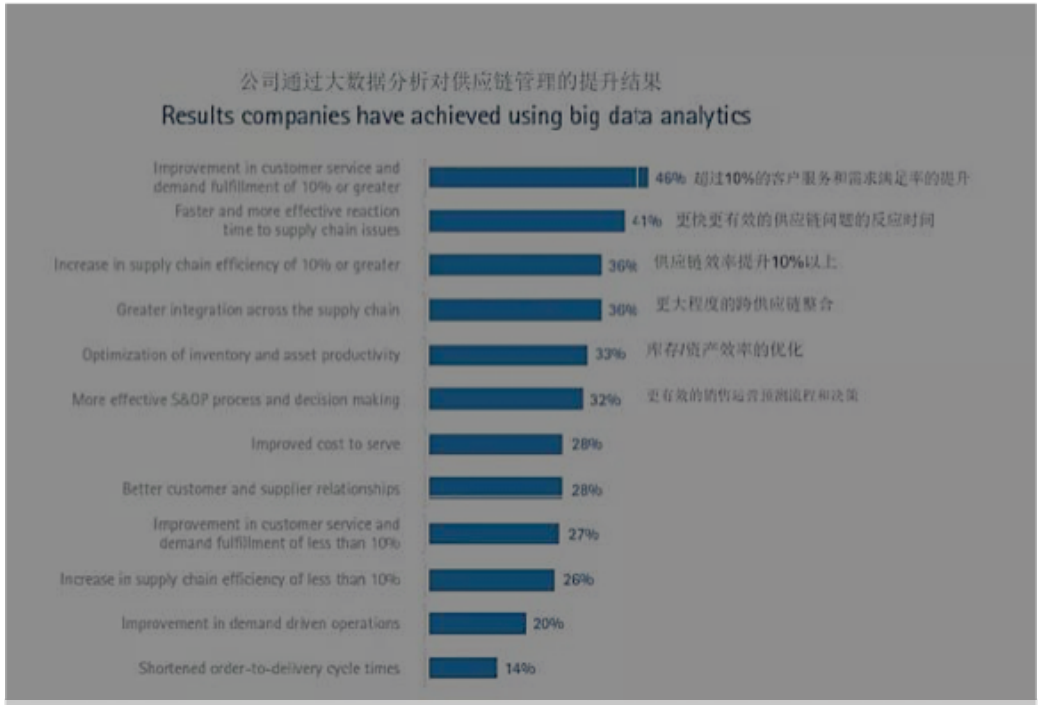


基于菜品标准化分类的口味/食材统计





客如云大数据服务带给B 端商户的价值

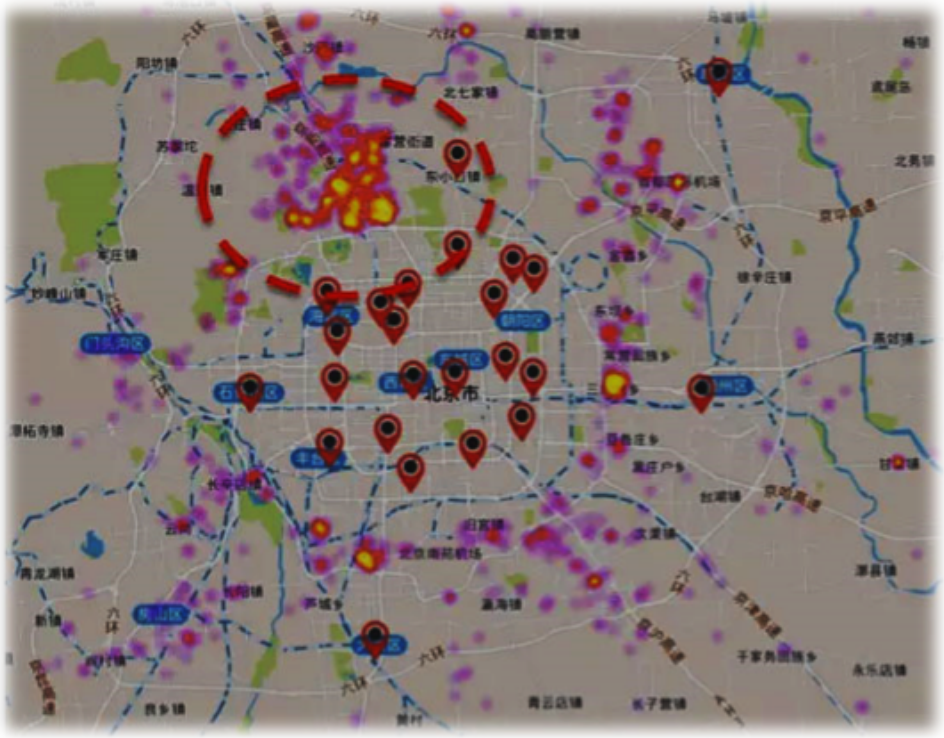


客如云大数据服务可以帮助企业将对供应链问题的反应时间提升41%；

将供应链效率提升10%甚至超过36%；

跨供应链的整合提升至30%以上。

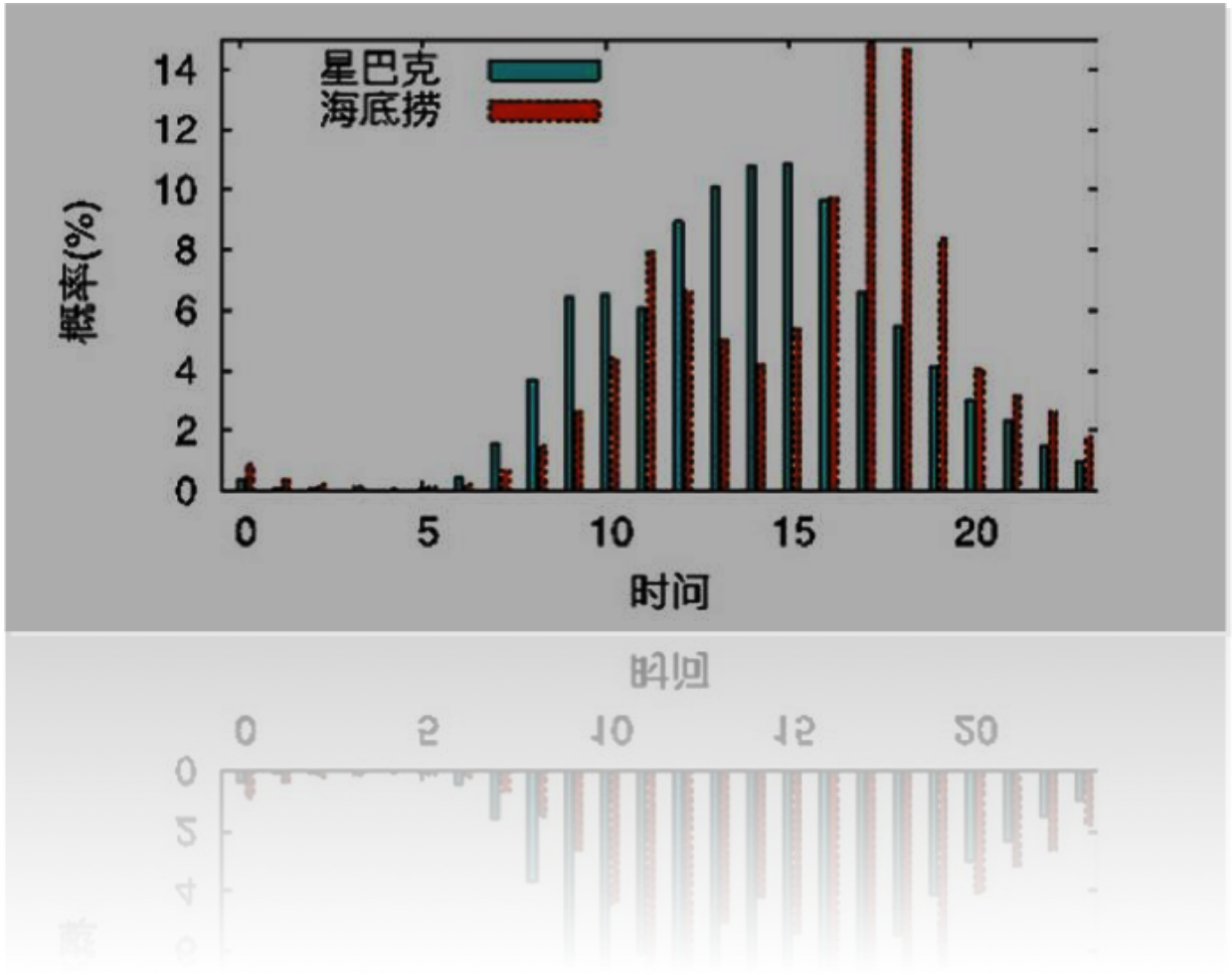
客如云大数据服务对B端商户供应链的价值



上图为网点未覆盖的用户需求分布热力，
红色的点表示现有某连锁餐饮品牌的门店位置。

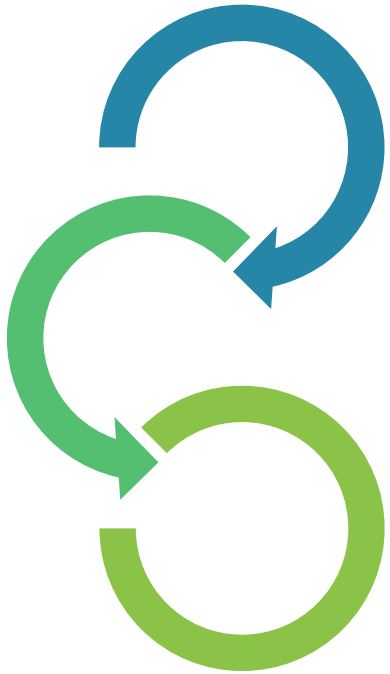
根据营业数据、客单价预测对开店选址给出精准的建议。

客如云大数据服务对B端商户商圈选择的价值



通过对顾客的一天内消费行为的数据分析，给出营销时间的专业建议。

客如云大数据服务对B端商户个性化服务的价值



基于店内员工工作情况的数据，
分析人效，优化店内运营效率

- 员工日均多休息1个小时；
- 翻台率提升7%；
- 员工流失率下降12%；
- 企业收益提升25%。

客如云大数据服务对B端商户人效提升的价值

钱包 QianBao.com

拥有**近亿**消费者以及消费者的行为和信用分析

拥有超过**上万家**供应商和供应商的销售数据与信用分析

通过数据，发掘客户需求，建立有效信用体系，
为商家与消费者提供贷款

客如云大数据服务对B端商户资金需求管理的价值

- 社交媒体、点餐平台
- 用户年龄、性别、职业、地域、行为轨迹、订单信息
- 调整营销策略、个性化推荐、营销组合设计、打造品牌形象、优化产品设计

锁定高消费能力的顾客，唤醒沉睡顾客，
丰富营销活动

客如云大数据服务对B端商户营销活动的价值

ArchData技术峰会
聚焦人工智能、大数据、基础架构、区块链等前沿课题

中生代技术提供咨询内训服务
技术架构，研发管理，敏捷开发，大数据微服务，AI，机器学习等

中生代技术提供人才服务
对接研发主管，内推精准人才