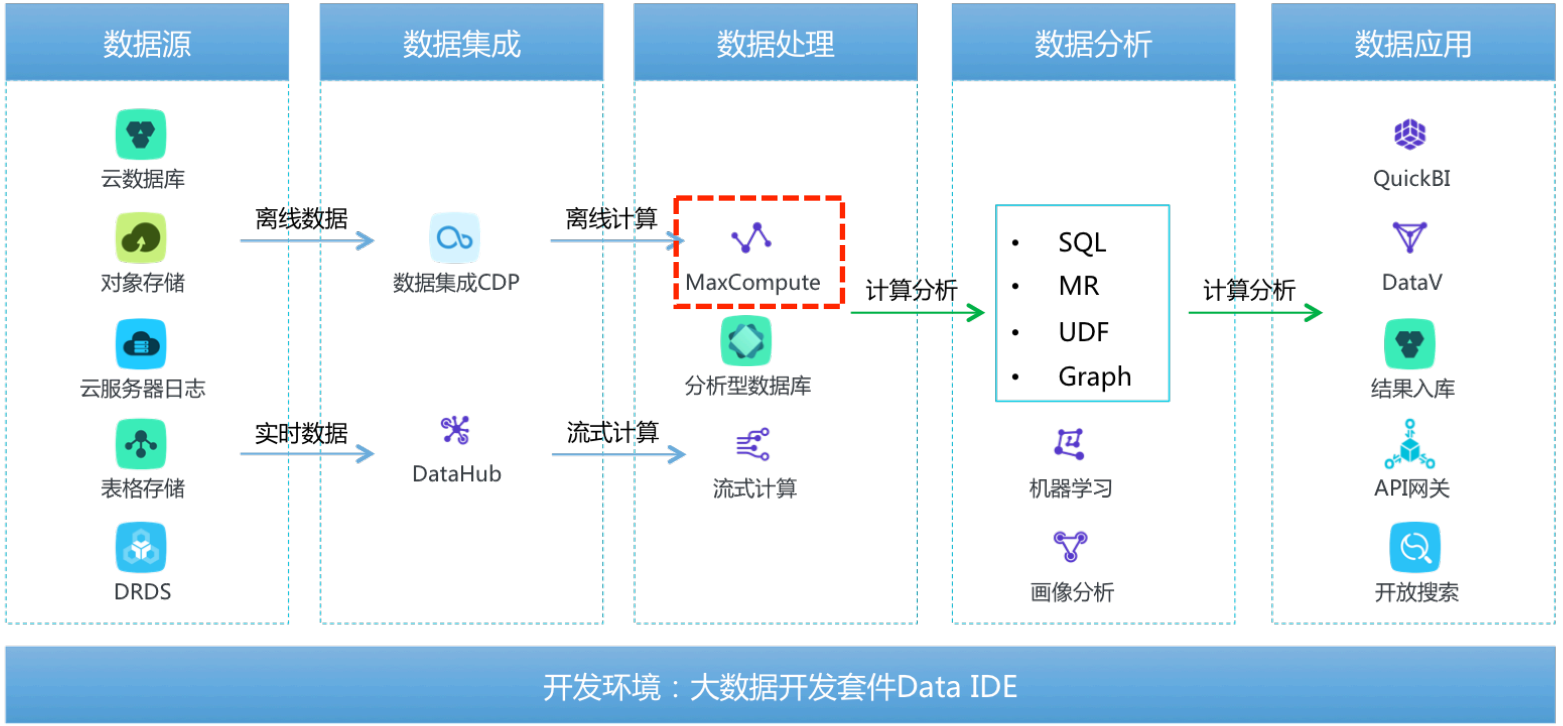
阿里云 MVP

如何使用大数据计算服务MaxCompute进行数据分析

企业云上搭建的数据分析平台



大数据计算服务MaxCompute的介绍

ODPS

大数据计算服务（MaxCompute，原ODPS）由阿里云自主研发，提供针对TB/PB级数据、实时性要求不高的分布式处理能力，应用于数据分析、挖掘、商业智能等领域。阿里巴巴的数据业务都运行在ODPS上。

分布式
采用分布式集群架构
跨集群技术突破
机群规模可以根据需要灵活扩展

安全
自动存储容错机制
所有计算在沙箱中运行
保障数据高安全性、高可靠性

易用
标准API的方式提供服务
高并发高吞吐量数据上传下载
全面支持基于SQL的数据处理

管理与授权
支持多用户管理协同分析数据
支持多种方式对用户权限管理
配置灵活的数据访问控制策略

大数据计算服务MaxCompute的特点

ODPS

大数据计算服务（MaxCompute，原ODPS）由阿里云自主研发，提供针对TB/PB级数据、实时性要求不高的分布式处理能力，应用于数据分析、挖掘、商业智能等领域。阿里巴巴的数据业务都运行在ODPS上。

海量运算触手可得

根据数据规模自动调整集群存储和计算能力，最大化发挥数据的价值

服务“开箱即用”

仅需简单的几步操作，就可以上传数据、分析数据并得到分析结果

数据存储安全可靠

三重备份、读写鉴权、应用沙箱、系统沙箱等多层次安全机制

多用户协作

保障数据安全的前提下最大化工作效率

按量付费

根据实际使用收费，最大化降低数据使用成本

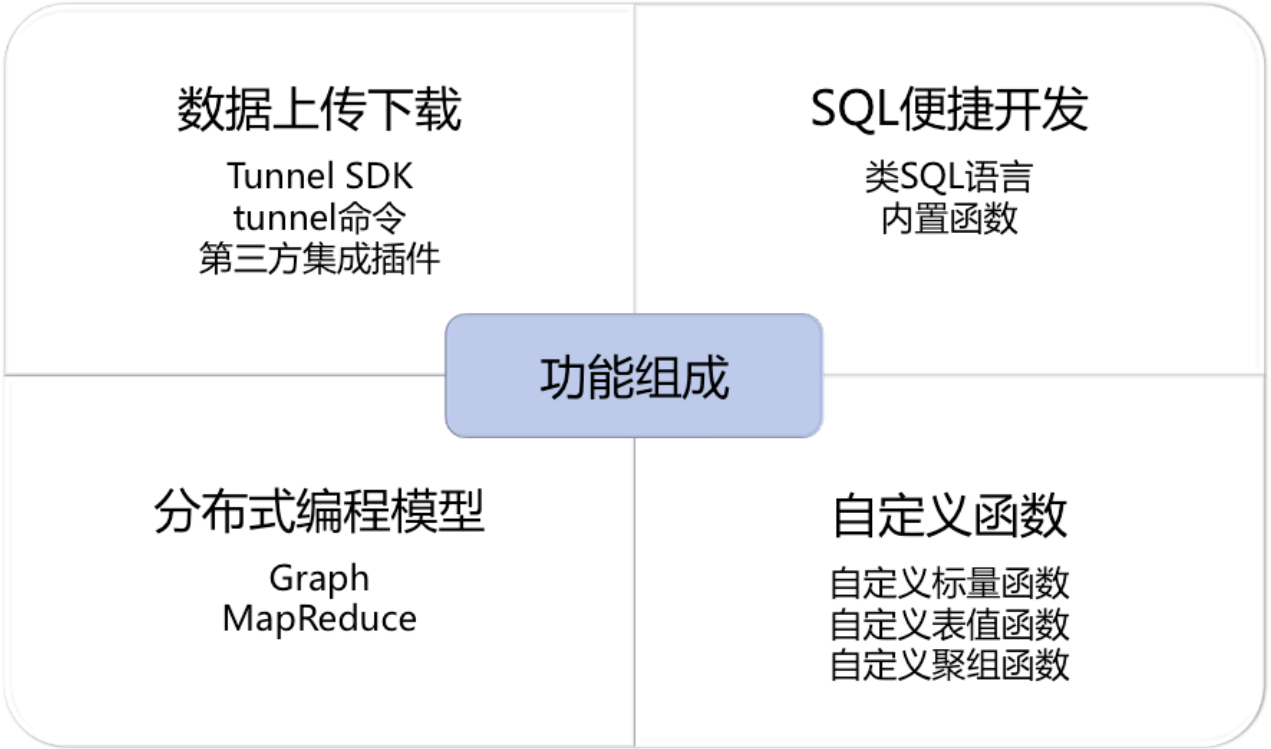
MaxCompute的使用场景

基于SQL构建大规模数据仓库系统和BI系统

基于DAG/Graph构建大型分布式应用系统

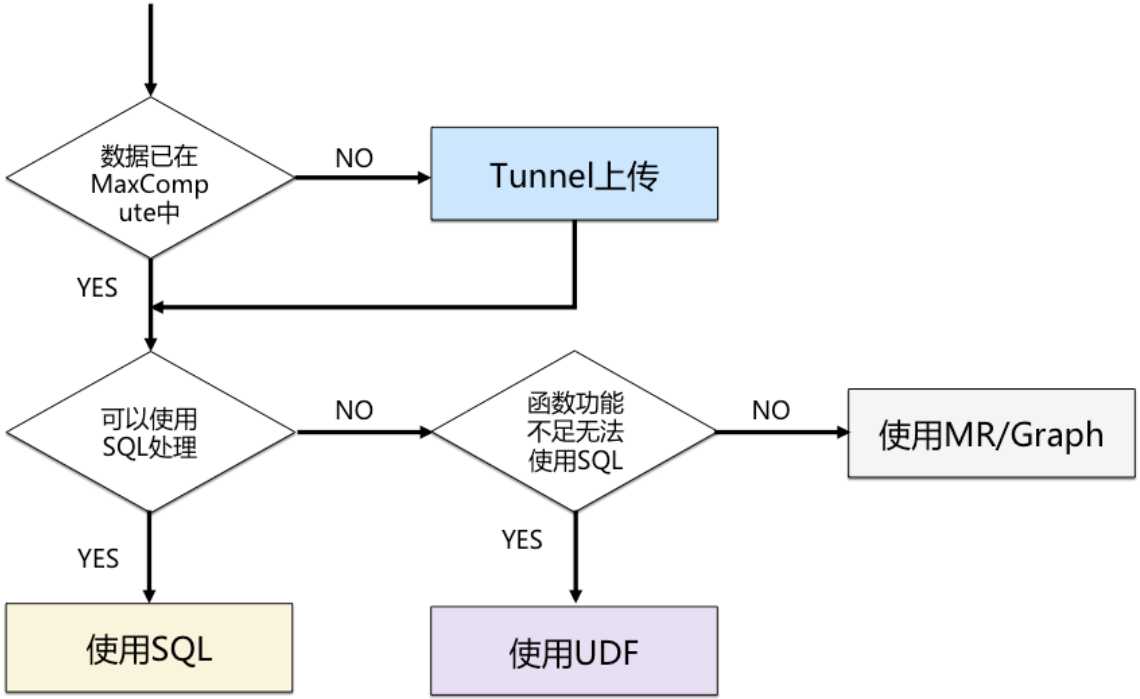
基于统计和机器学习的大数据统计和数据挖掘

MaxCompute的功能组成

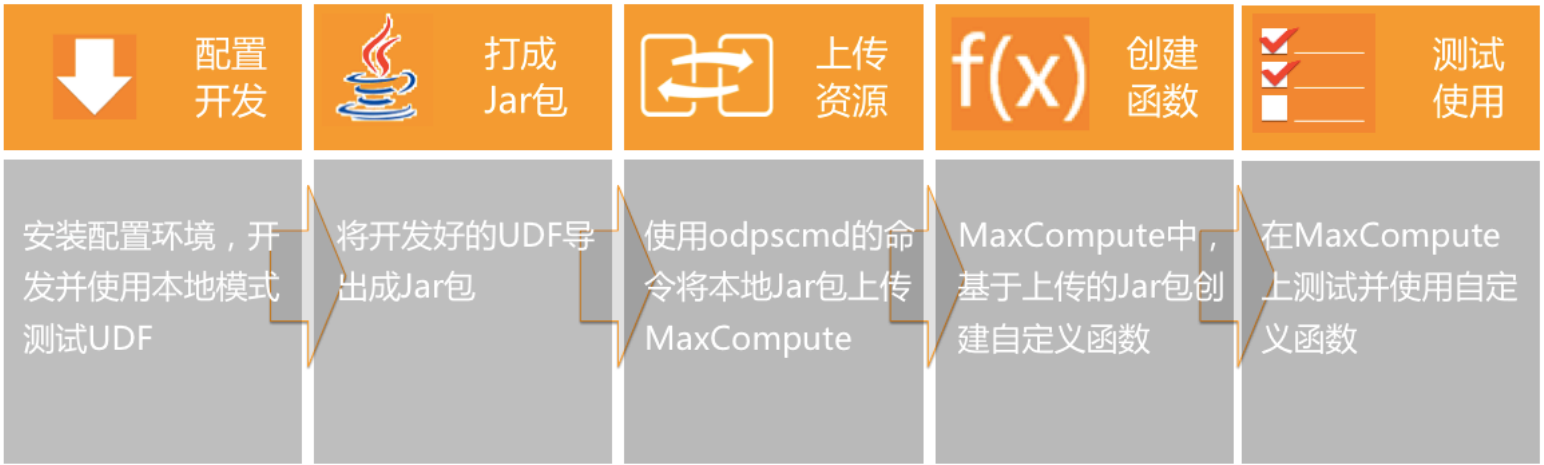


MaxCompute的功能选择

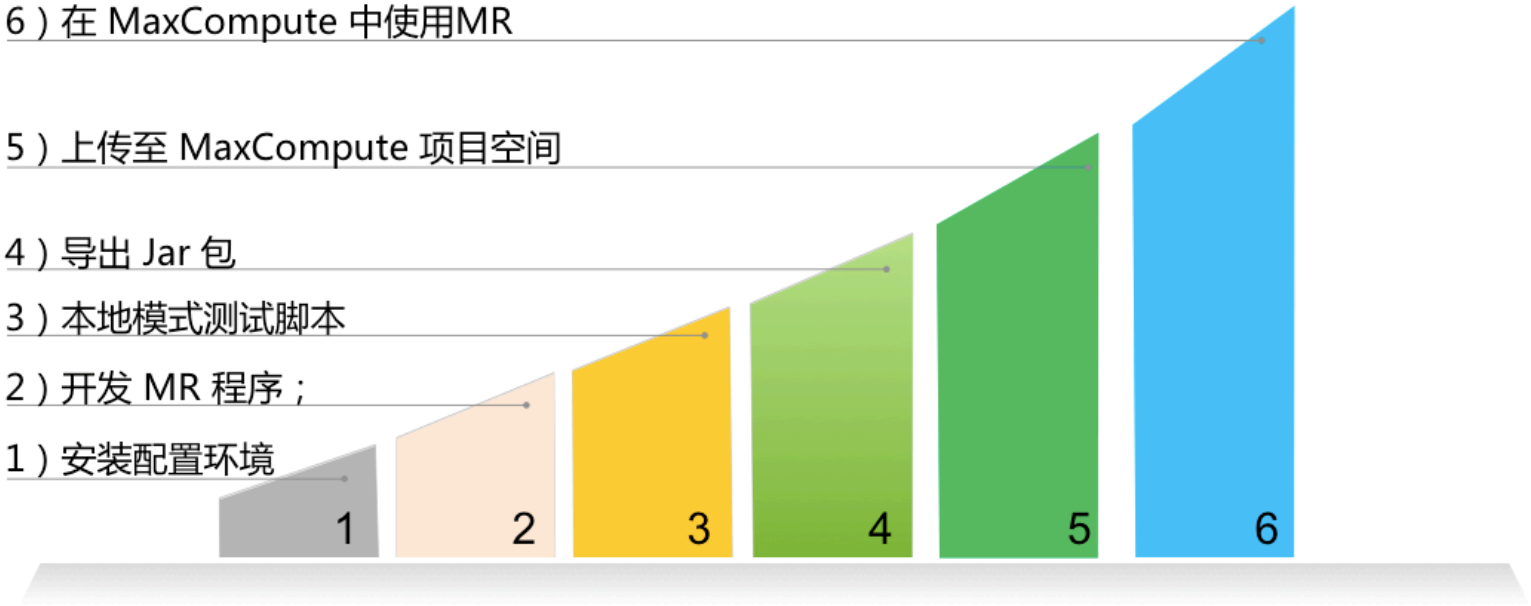
原则： 能使用SQL的情形下，尽量使用SQL



自定义函数UDF的开发流程



MR的开发流程


6）在 MaxCompute 中使用MR
5）上传至 MaxCompute 项目空间
4）导出 Jar 包
3）本地模式测试脚本
2）开发 MR 程序；
1）安装配置环境
1
2
3
4
5
6

使用MaxCompute的注意事项：元数据不一致

常用方法： 定义标准元数据，创建Map表或者定义转换规则，清洗数据

customers

Id	Gender	Age	Height
1	M	36	176
2	F	27	162
3	F	49	174
4	M	16	165

kh

BM	XB	HF	SG
30001	女	YH	5'3"
2987	男	YH	5'9"
9527	男	WH	5'4"
101	女	YH	5'7"

map_gender

Src_value	Std_id	Src_op
M	1	Ebuzi
F	2	Ebuzi
女	2	Service
男	1	Service

身高：厘米

公式：round((x*12+y)*2.54,0)

ods_customers

Id	Gender	Age	Height
1	1	36	176
2	2	27	162
3	2	49	174
4	1	16	165

ods_kh

BM	XB	HF	SG
30001	2	YH	160
2987	1	YH	175
9527	1	WH	163
101	2	YH	170

使用MaxCompute的注意事项：数据缺失

常用方法： 1-填充固定值；2-填充统计值；3-填充拟合值；

ods_customers

Id	Gender	Age	Height
1	1	36	176
2	2	27	162
3	2	49	174
4	1	16	165
5	1	23	180
6	NULL	31	159
.....	.....	.....	.....

① 填充固定值，NULL值统一由 -1 代替，代表未知

Id	Gender	Age	Height
6	-1	31	159

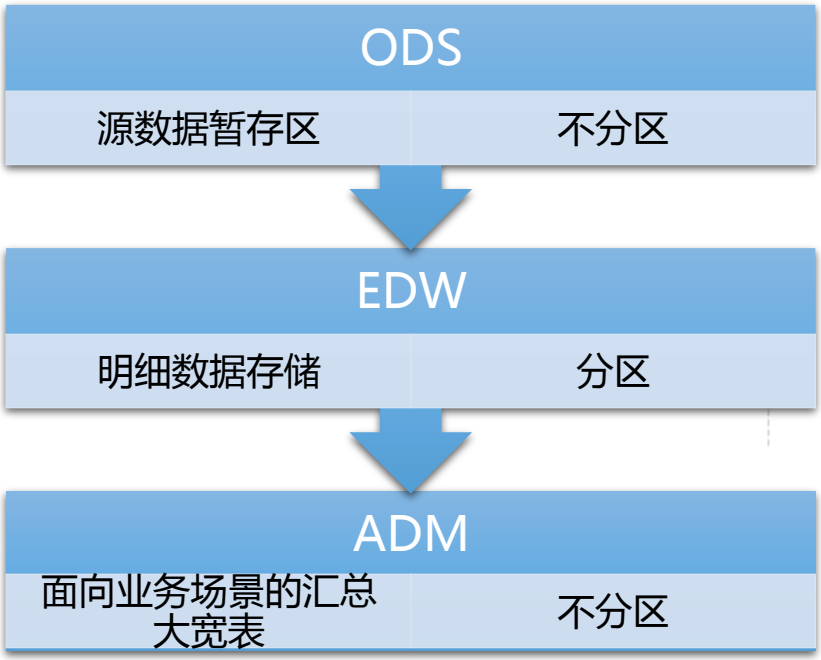
② 填充统计值，如均值、极值、众数等

Id	Gender	Age	Height
6	1	31	159

③ 通过模型或者规则拟合，得到“最应该”填的值：

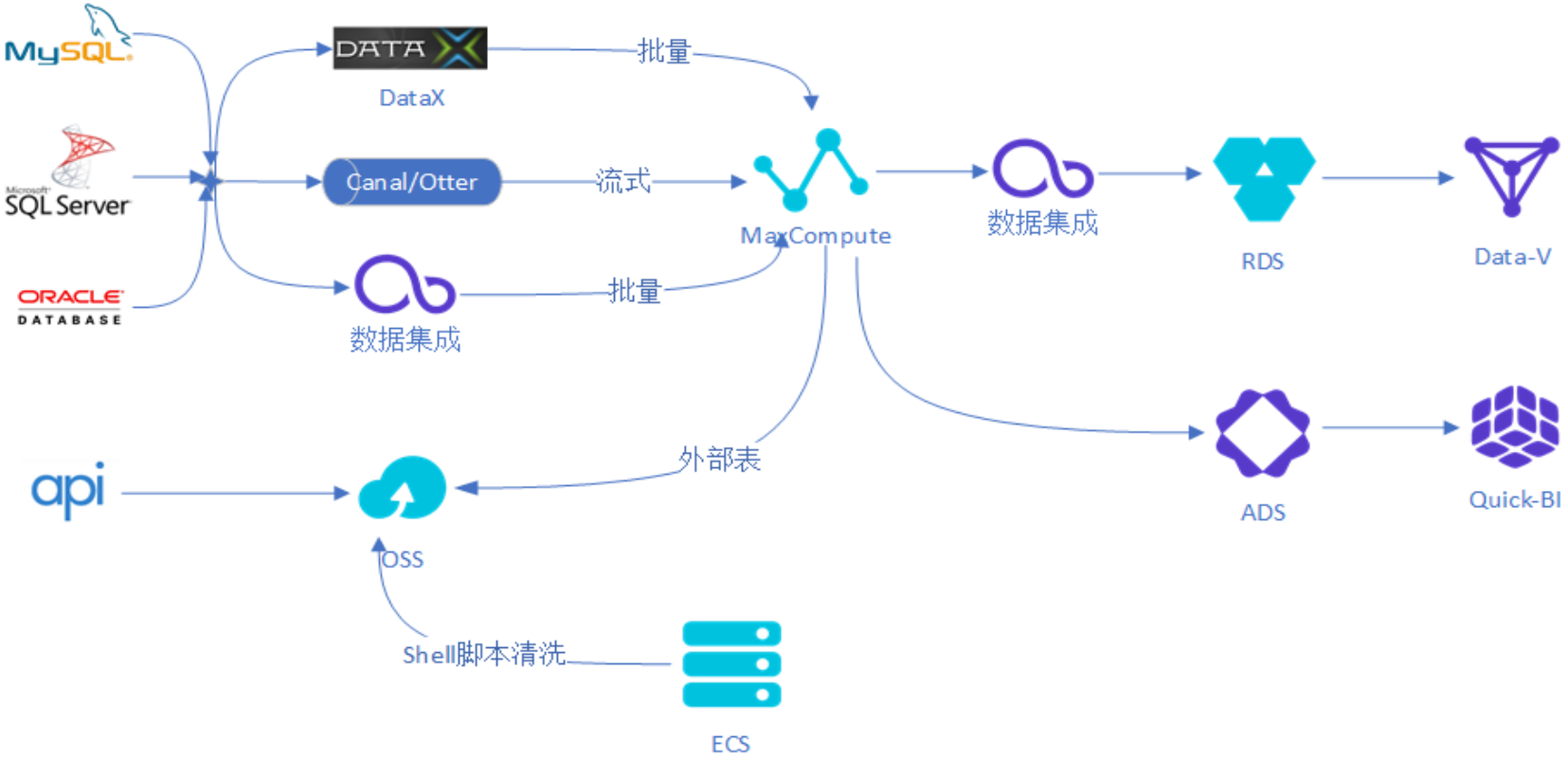
Id	Gender	Age	Height
6	2	31	159

使用MaxCompute的注意事项：数仓的搭建



- ODS实现准实时、跨系统的运营细节数据的查询，以获得细粒度的运营数据展现；
- EDW实现基于历史数据的统计分析和数据挖掘，以获得客户深层次的特征和市场发展的规律；
- ADM在基础数据的上进行加工汇总形成的指标数据存储分析型和加工汇总型数据。

驻云的客户案例：某智慧商场的大数据架构



Thanks!