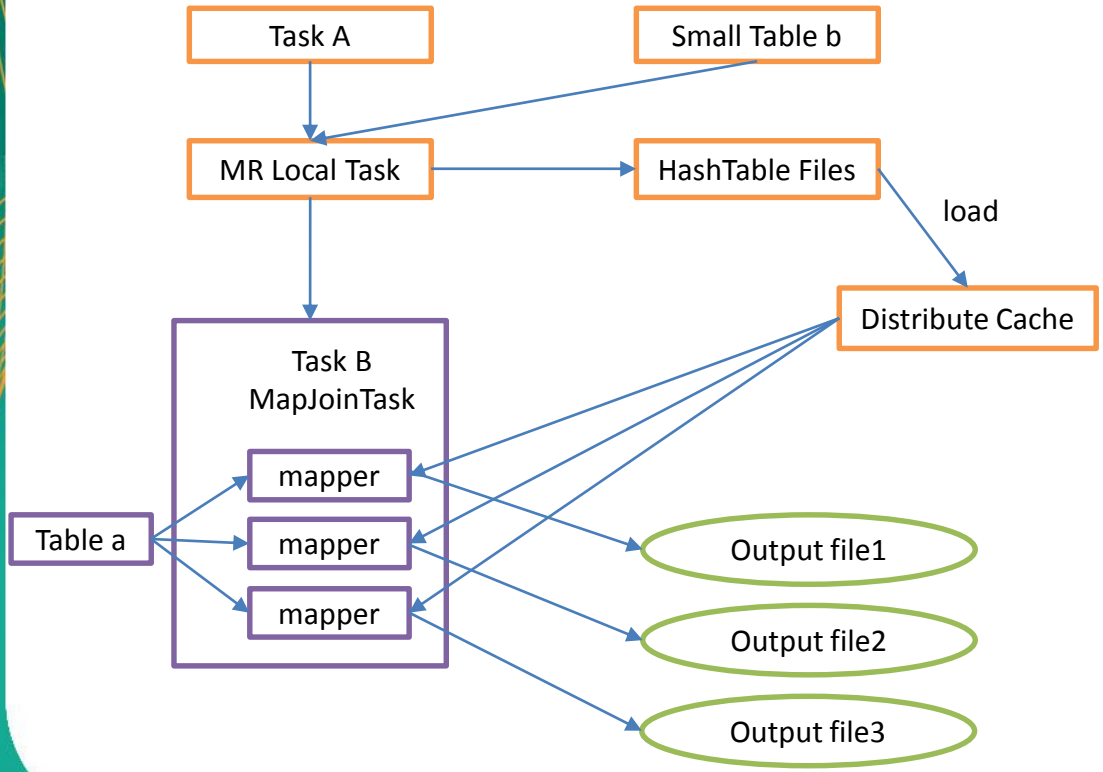
大数据技术图解

签名：汪辉

MapJoin

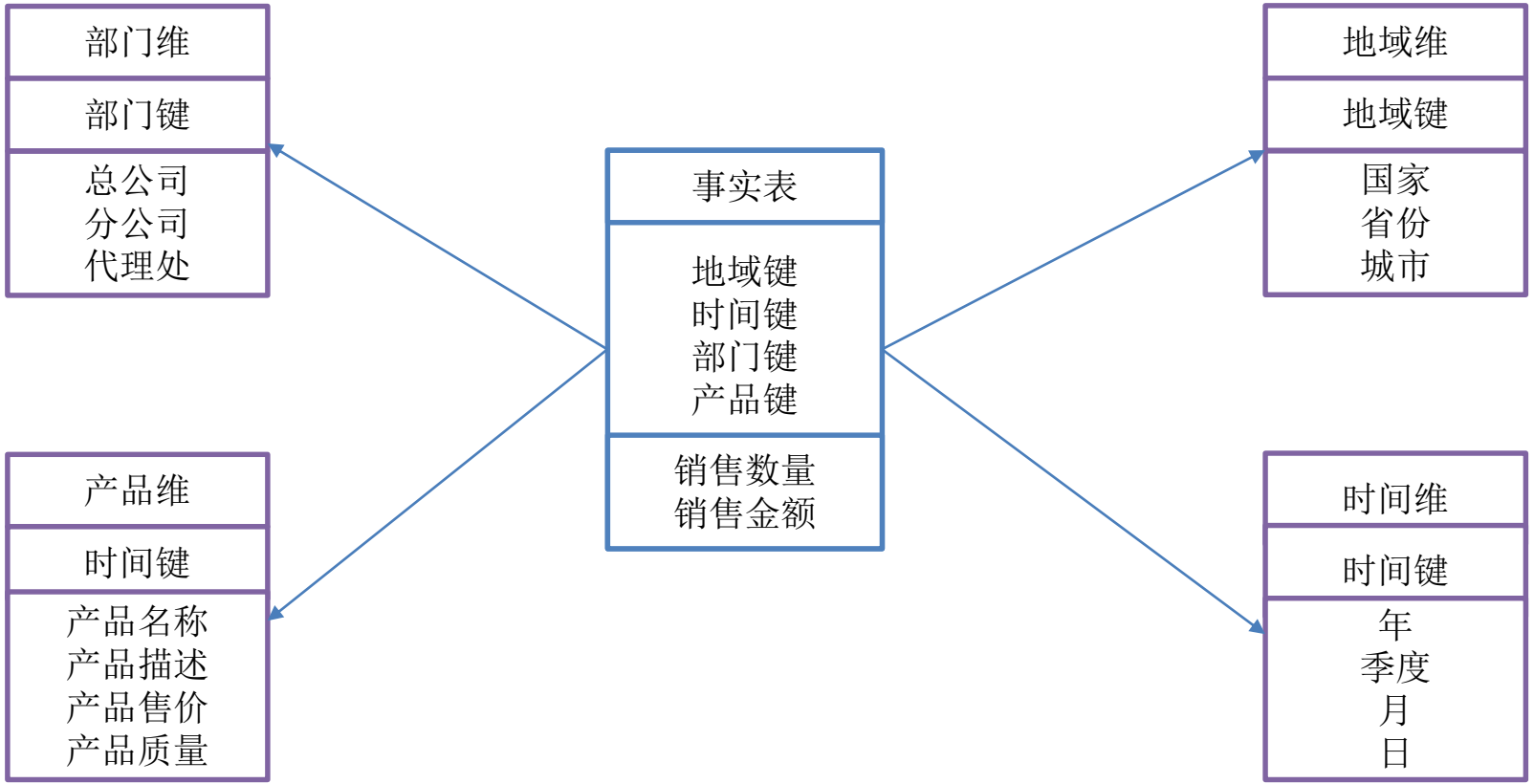


1）Task A，它是一个Local Task（在客户端本地执行的Task），负责扫描小表b的数据，将其转换成一个HashTable的数据结构，并写入本地的文件中，之后将该文件加载到DistributeCache中。

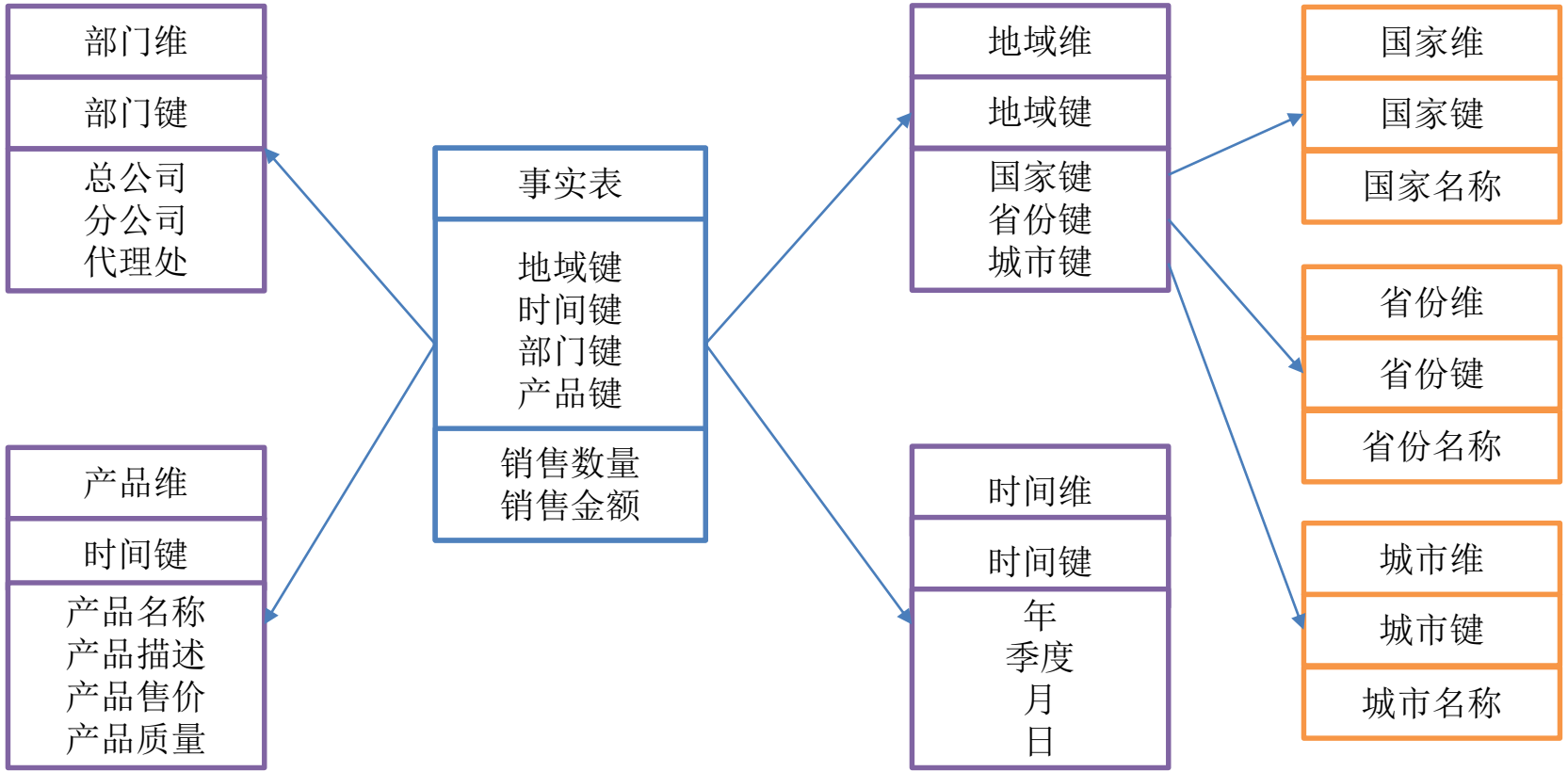
2）Task B，该任务是一个没有Reduce的MR，启动MapTasks扫描大表a，在Map阶段，根据a的每一条记录去和DistributeCache中b表对应的HashTable关联，并直接输出结果。

3）由于MapJoin没有Reduce，所以由Map直接输出结果文件，有多少个Map Task，就有多少个结果文件。

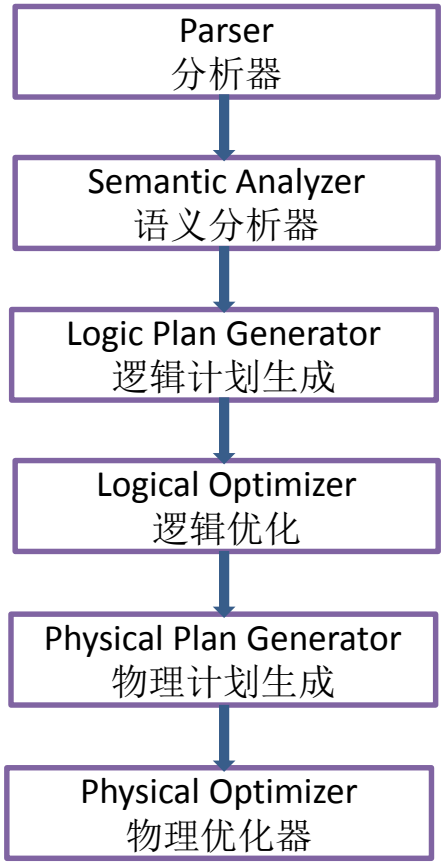
星型模型


部门维
部门键
总公司
分公司
代理处
地域维
地域键
国家
省份
城市
事实表
地域键
时间键
部门键
产品键
销售数量
销售金额
产品维
时间键
产品名称
产品描述
产品售价
产品质量
时间维
时间键
年
季度
月
日

雪花模型


部门维
部门键
总公司
分公司
代理处
产品维
时间键
产品名称
产品描述
产品售价
产品质量
事实表
地域键
时间键
部门键
产品键
销售数量
销售金额
地域维
地域键
国家键
省份键
城市键
时间维
时间键
年
季度
月
日
国家维
国家键
国家名称
省份维
省份键
省份名称
城市维
城市键
城市名称

Hive编译器组成



将HQL转换成抽象语法树

将抽象语法树换成查询块

将查询块转换成逻辑查询计划

重写逻辑查询计划

将逻辑计划转成物理计划MR jobs

选择最佳的策略


尚硅谷
www.atguigu.com