# 基于阿里云搭建数据仓库（离线）

阿里云大学 & 尚硅谷  联合出品

# 课程目标

1）学习**搭建**一个**数据仓库**的过程，理解数据在整个数仓架构的从**采集、存储、计算、输出、展示**的整个业务流程。

2）整个数仓体系完全搭建在**阿里云架构**上，理解并学会**运用各个服务组件**，了解各个组件之间如何**配合联动**。

3）前置知识要求

- 熟练掌握**SQL语法**
- 熟悉**Linux**命令
- 对**Hadoop大数据体系**有一定的了解

# 第1章 课程目录

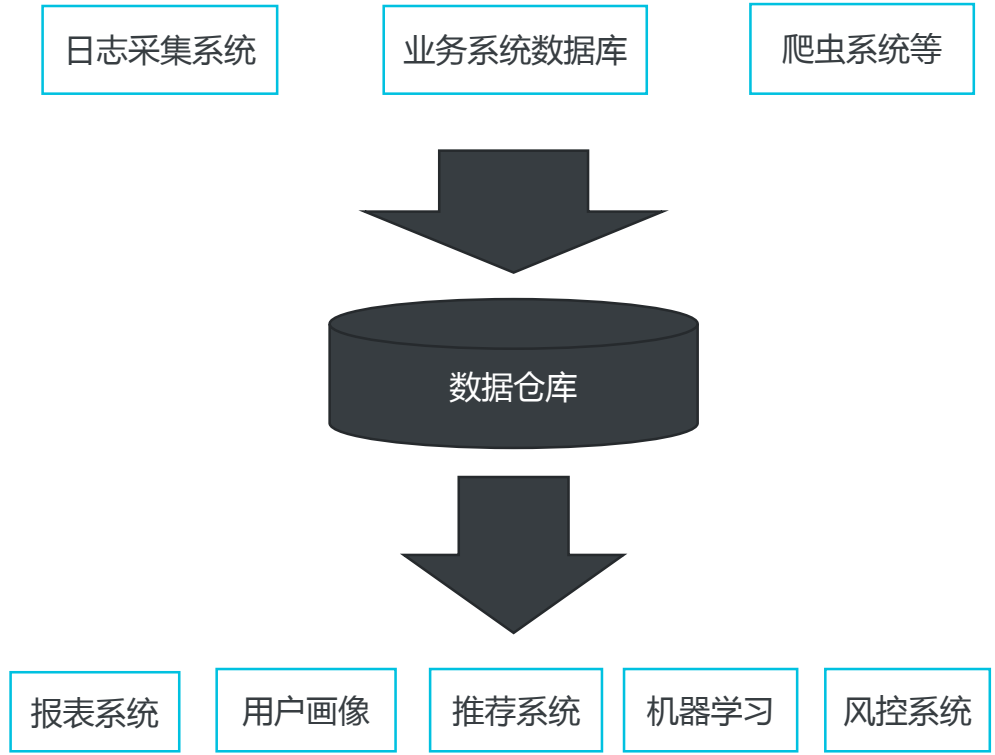
# 第1章 数据仓库概念

数据仓库定义（ Data Warehouse ），是为企业所有决策制定过程，提供所有系统数据支持的战略集合。

数据仓库好处：可以帮助企业，改进业务流程、控制成本、提高产品质量等。

数据仓库做什么：清洗，转义，分类，重组，合并，拆分，统计等。

数据仓库输出到哪：报表系统、用户画像、推荐系统、机器学习、风控系统等



# 第2章 课程目录

# 2.1 项目需求分析

1）采集埋点日志数据

2）采集业务数据库中数据

3）数据仓库的搭建（用户行为数仓、业务数仓）

4）分析统计业务指标

5）对结果进行可视化展示

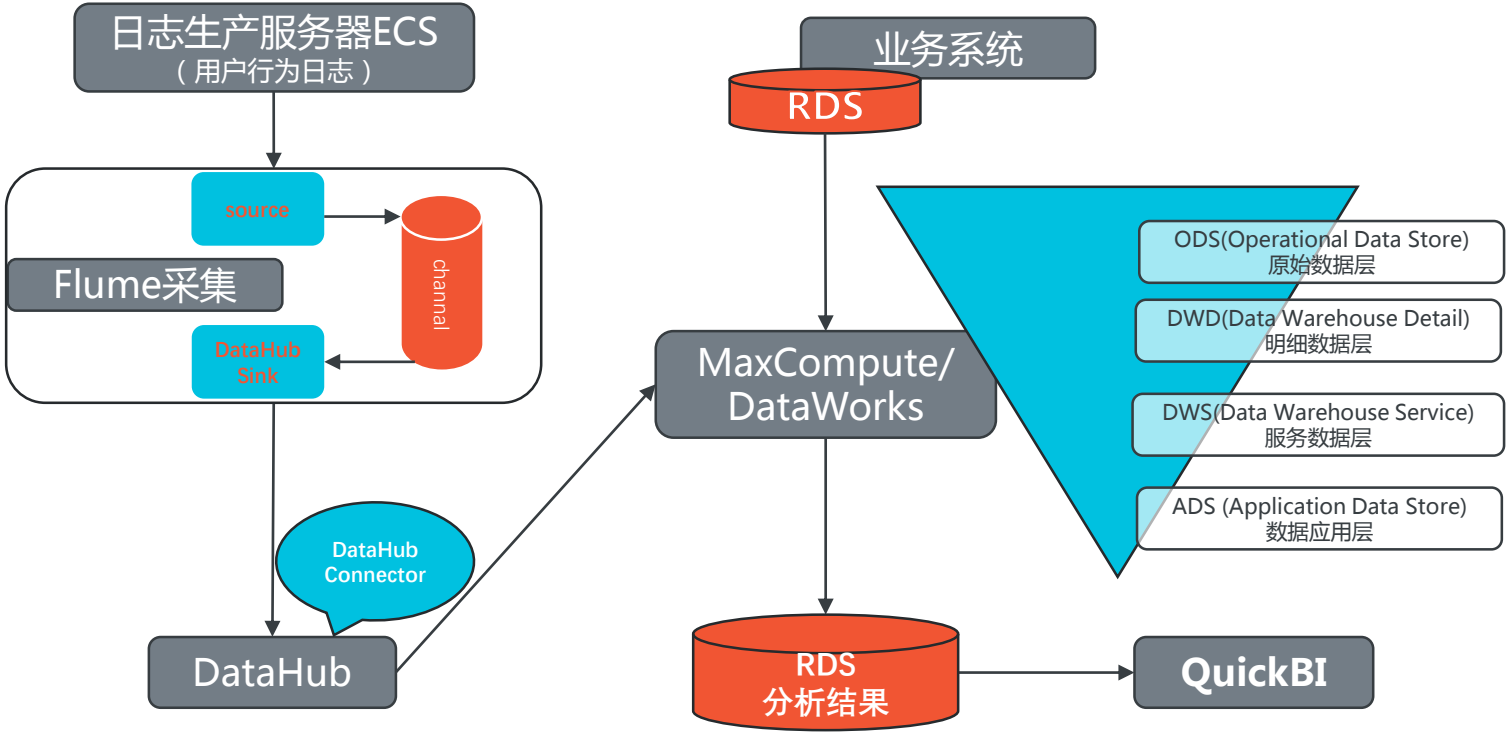
# 2.2 阿里云技术框架

| 阿里云产品 | 简介 | 类比 |
| --- | --- | --- |
| DataHub | 数据总线 | Kafka +各种服务接口 |
| MaxCompute | 大数据计算框架 | Hadoop+Hive+调度器 |
| DataWorks | 可视化MaxCompute的开发管理平台 | 目前没有 |
| RDS | 关系型数据库 | MySql |
| QuickBI | 可视化数据展示工具 | Tableau、Echarts、Kibana |
| ECS | 弹性服务器 | Linux服务器 |

# 2.2.1 技术选型

| | 阿里云框架 | 开源框架 |
| --- | --- | --- |
| ➢ 数据采集传输： | Flume、DataHub、RDS | Flume、Kafka、Sqoop 、DataX |
| ➢ 数据存储： | MaxCompute、DataWorks | MySql、Hadoop、HBase |
| ➢ 数据计算： | MaxCompute、DataWorks | Hive、Spark、Flink |
| ➢ 数据可视化： | QuickBI | Tableau、Echarts、Kibana |

# 2.2.2 系统数据流程设计



# 2.2.3 服务器选型

## 服务器选择物理机还是云主机？

**1）机器成本考虑：**

物理机：以128G内存，20核物理CPU，40线程，8THDD和2TSSD硬盘，戴尔品牌单台报价4W出头，需考虑托管服务器费用。一般物理机寿命5年左右。

云主机：以阿里云为例，差不多相同配置，每年5W。

**2）运维成本考虑：**

物理机：需要有专业的运维人员，平均每月15000元；

云主机：很多运维工作都由阿里云完成，运维相对较轻松。

# 2.2.4 集群规模

## 1）用户行为数据

（1）每天日活跃用户100万，每人一天平均100条：100万*100条=10000万条

（2）每条日志1K左右，每天1亿条：100000000 / 1024 / 1024 = 约100G

（3）数仓ODS层采用LZO+parquet存储：100g压缩为10g左右

（4）数仓DWD层采用LZO+parquet存储：10g左右

（5）数仓DWS层轻度聚合存储（为了快速运算，不压缩）：50g左右

（6）数仓ADS层数据量很小：忽略不计

（7）保存3副本：70g*3=210g

（8）半年内不扩容服务器来算：210g*180天=约37T

（9）预留20%~30%Buf=37T/0.7=53T

## 2）DataHub中数据

（1）每天约100G数据*副本（2）=200g

（2）保存3天*200g=600g

（3）预留30%buf=600g/0.7=857g=约1T

## 3）Flume中默认缓存的数据比较小：暂时忽略不计

## 4）业务数据

（1）每天活跃用户100万，每天下单的用户10万，每人每天产生的业务数据10条，每条日志1k左右：10万*10条*1k=1g左右

（2）数仓四层存储：1g*3=3g

（3）保存3副本：3g*3=9g

（4）半年内不扩容服务器来算：9g*180天=约1.6T

（5）预留20%~30%Buf=1.6T/0.7=2T

## 5）集群总规模：53T+1T+2T=56T

## 6）算到这：约8T*7台服务器

# 2.2.5 购买服务器建议

以日均100G（日志+数据）为例

| 购买服务 | 建议配置 | 年成本 | 备注 |
| --- | --- | --- | --- |
| DataHub | medium | 目前免费中 | medium: 5000record/s |
| MaxCompute | 32CU*7 | 342720.00 | 1CU=1cpu+4G内存 |
| RDS | 4核8G | 10,914.00 | 存放离线统计结果 |
| QuickBI | 高级版(企业) | 38,207.00 | |
| 年总成本 | | 391841.41 | |
| 月均成本 | | 32653.42 | |

# 课程说明

第3~10章，采用Word课件授课


阿里云