数据处理平台之演进

姚延栋
yyao@pivotal.io

Pivotal

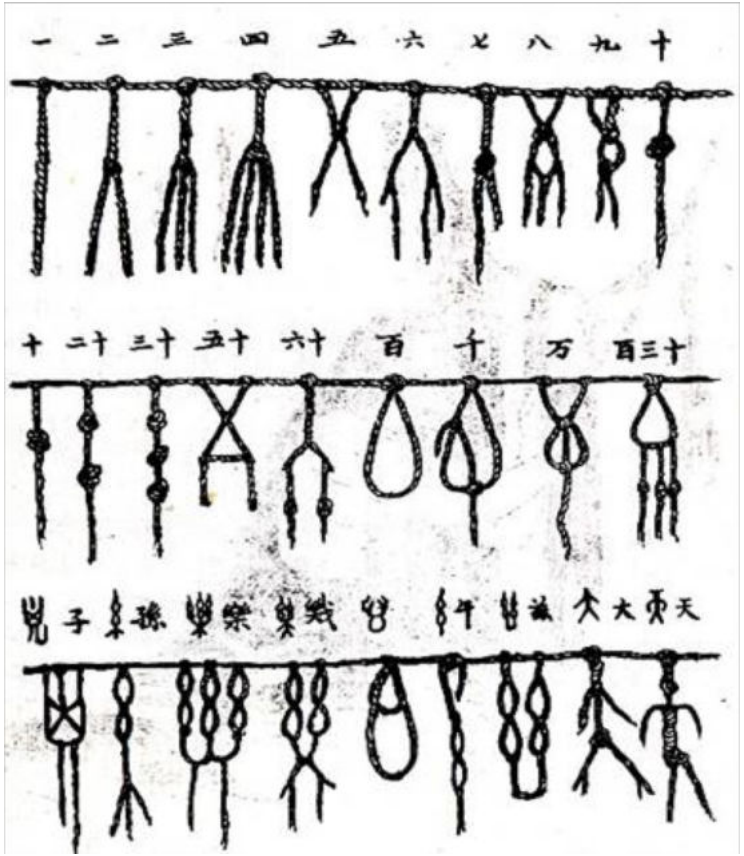
内容

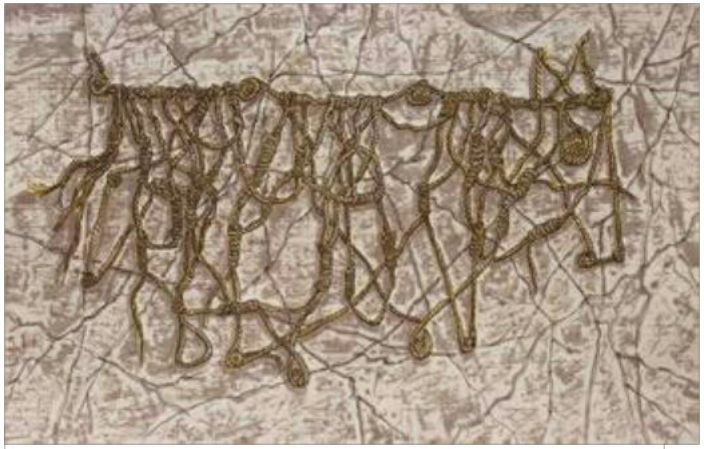
- 人工数据处理
- 半人工数据处理
- 早期电子数据处理
- 数据库
- NoSQL 数据库
- SQL 数据库回归
- 集成平台
- Greenplum

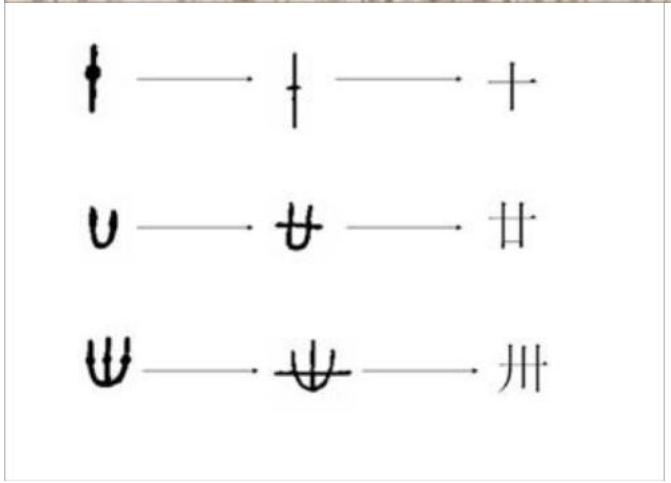
人工数据处理

Pivotal

上古时期 - 结绳记事



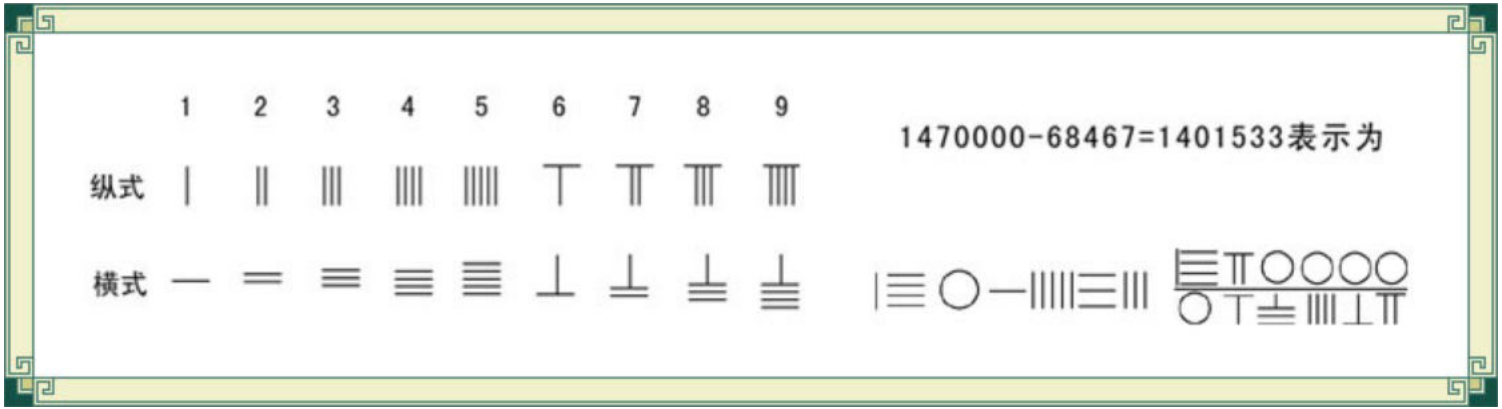




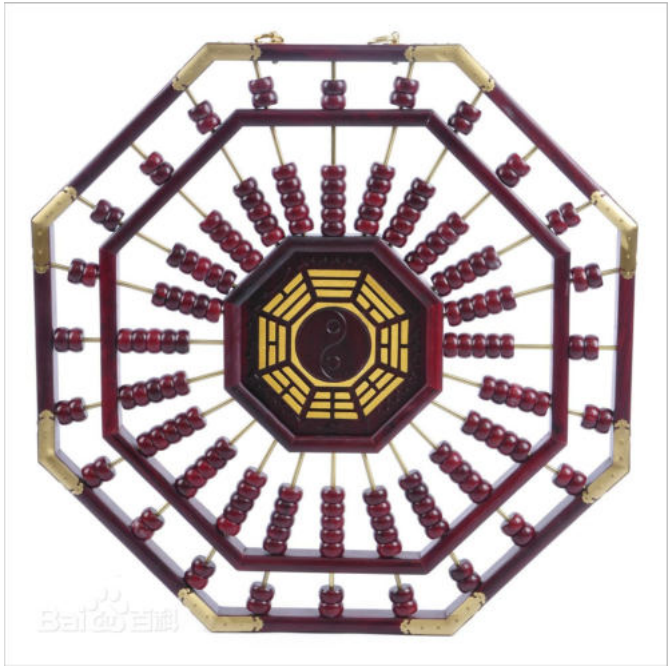
Pivotal

三皇五帝 - 书契

商周 - 算筹



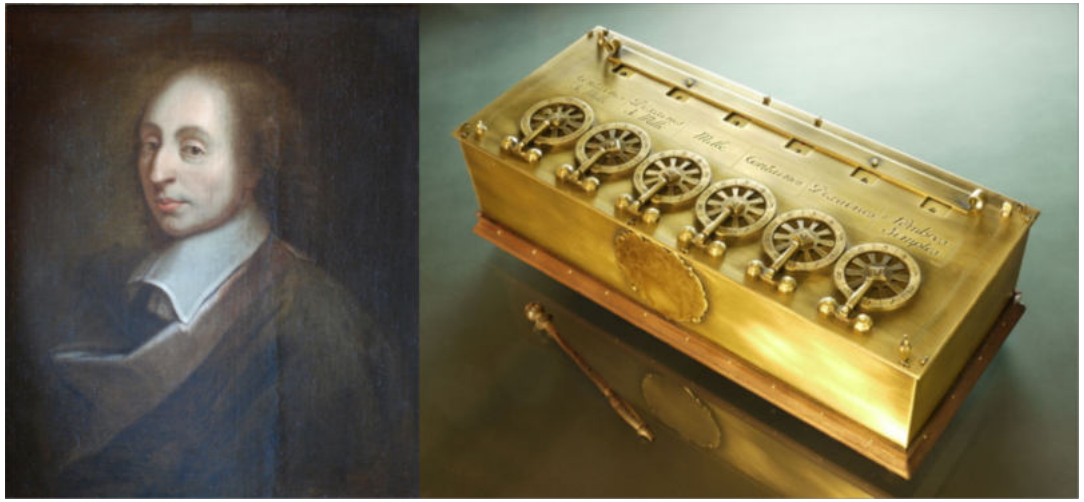
始于汉，兴于宋（960-1279）- 算盘

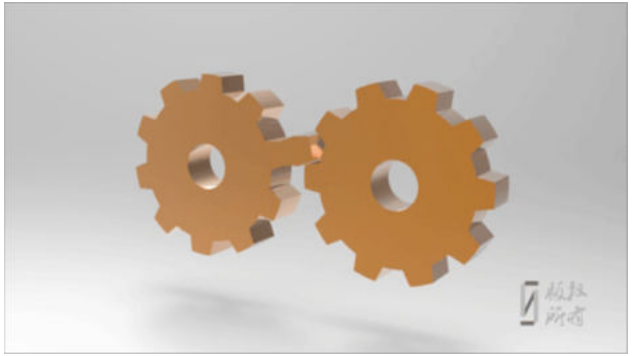


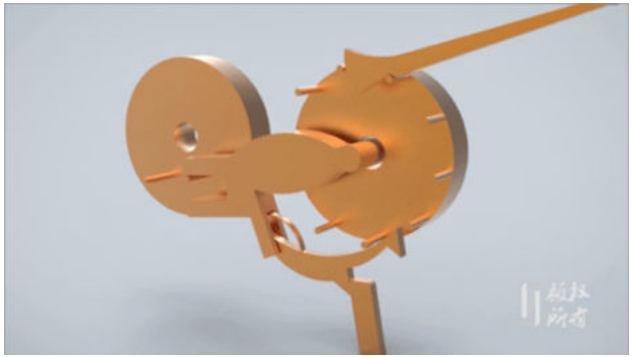
半人工数据处理

Pivotal

1640s - 帕斯卡加法器, 第一台机械计算机







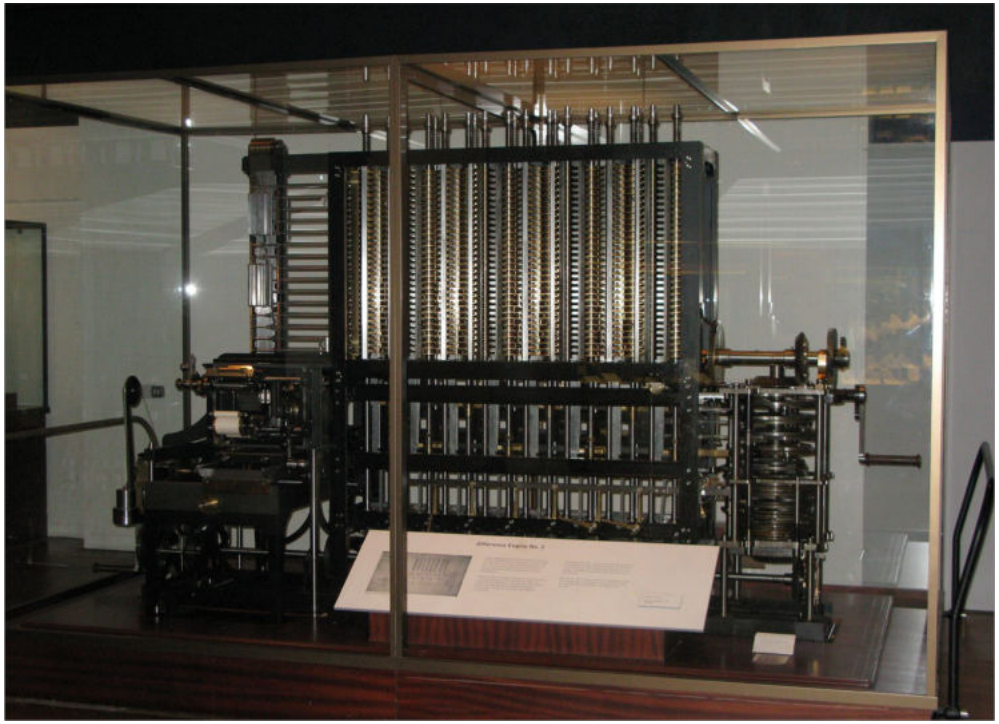
1670s - 莱布尼兹乘法器

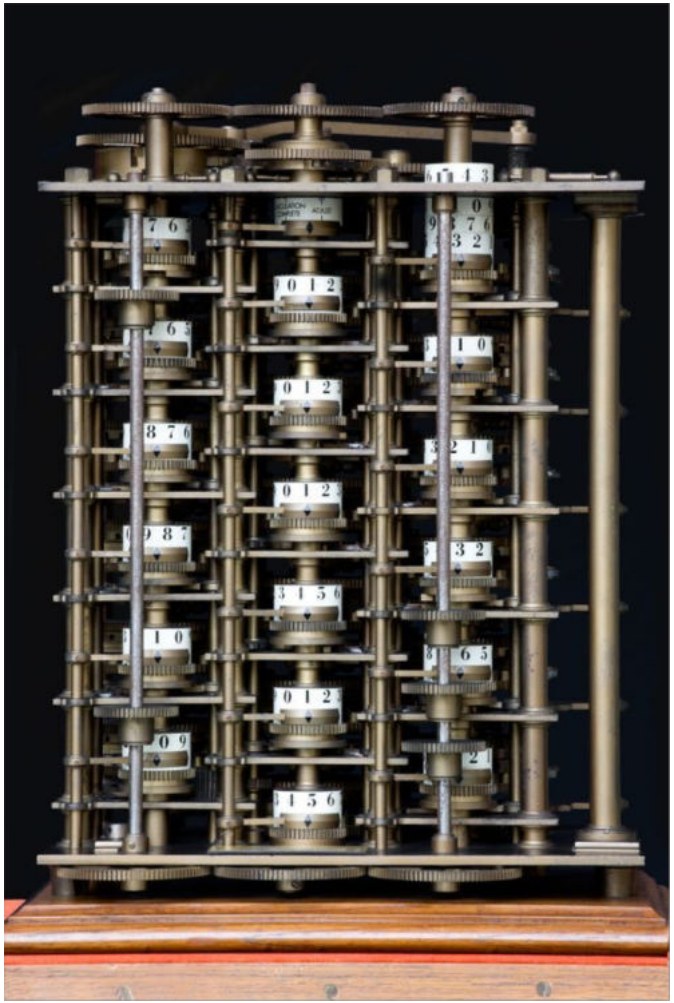
Pivotal

1805: 杰卡德提花编织机，指令程式化

Pivotal

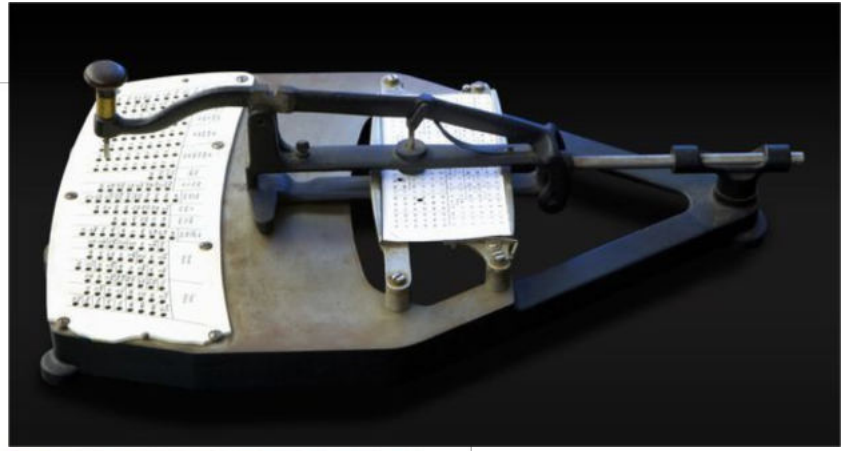
1822 - 巴贝奇差分机





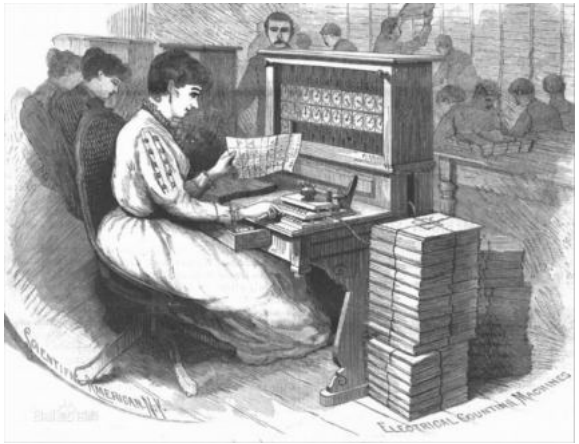
Pivotal

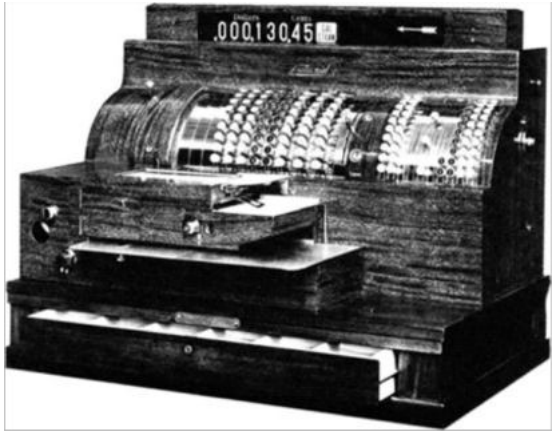
1888: 霍尔瑞斯制表机，数据存储和处理



Pivotal

1920s: 单元记录设备、电算机、制表机



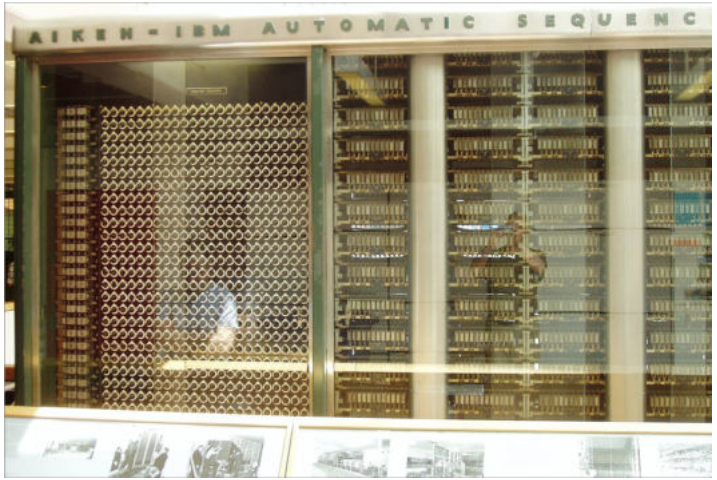


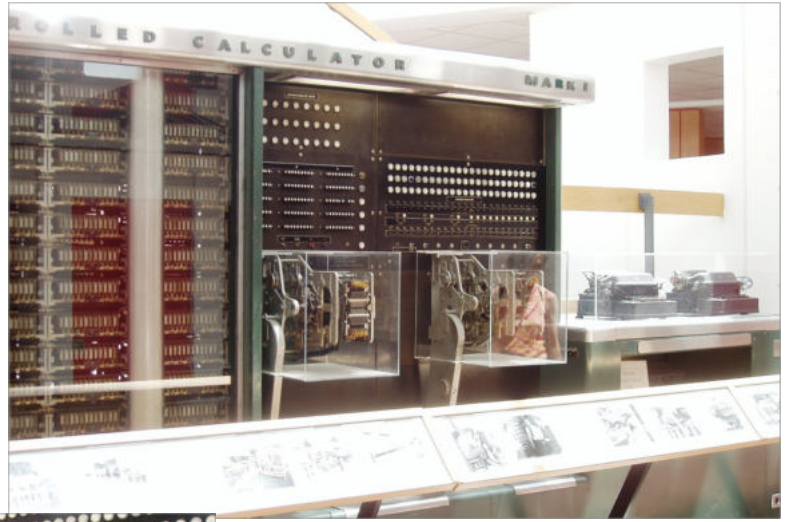
Pivotal

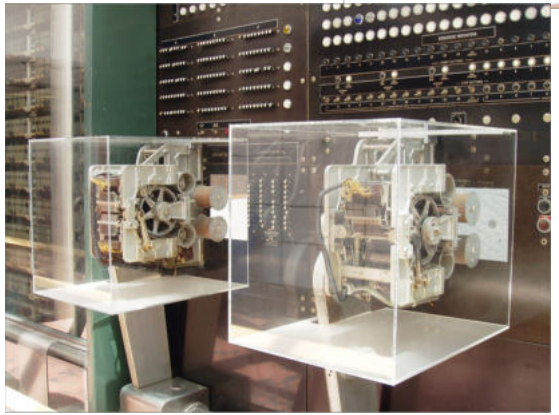
早期电子数据处理
电子计算机实现第一次整合

Pivotal

1944: MARK-1 早期计算机的最后代表

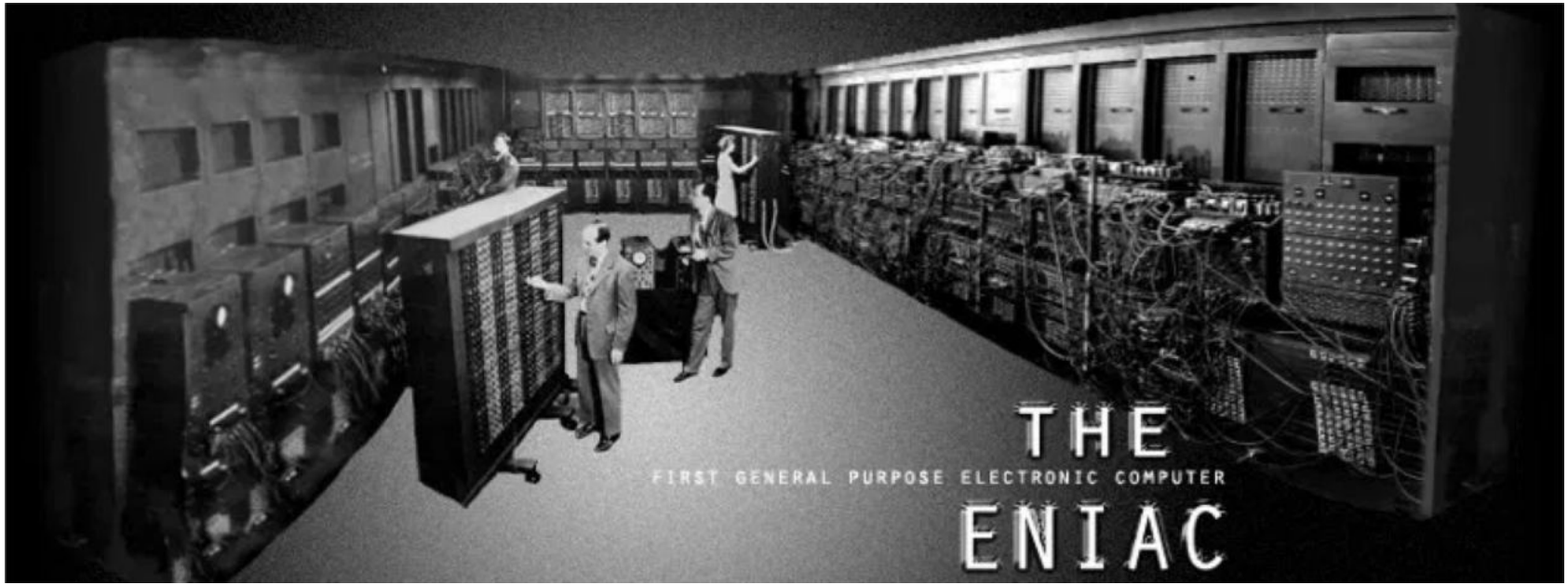




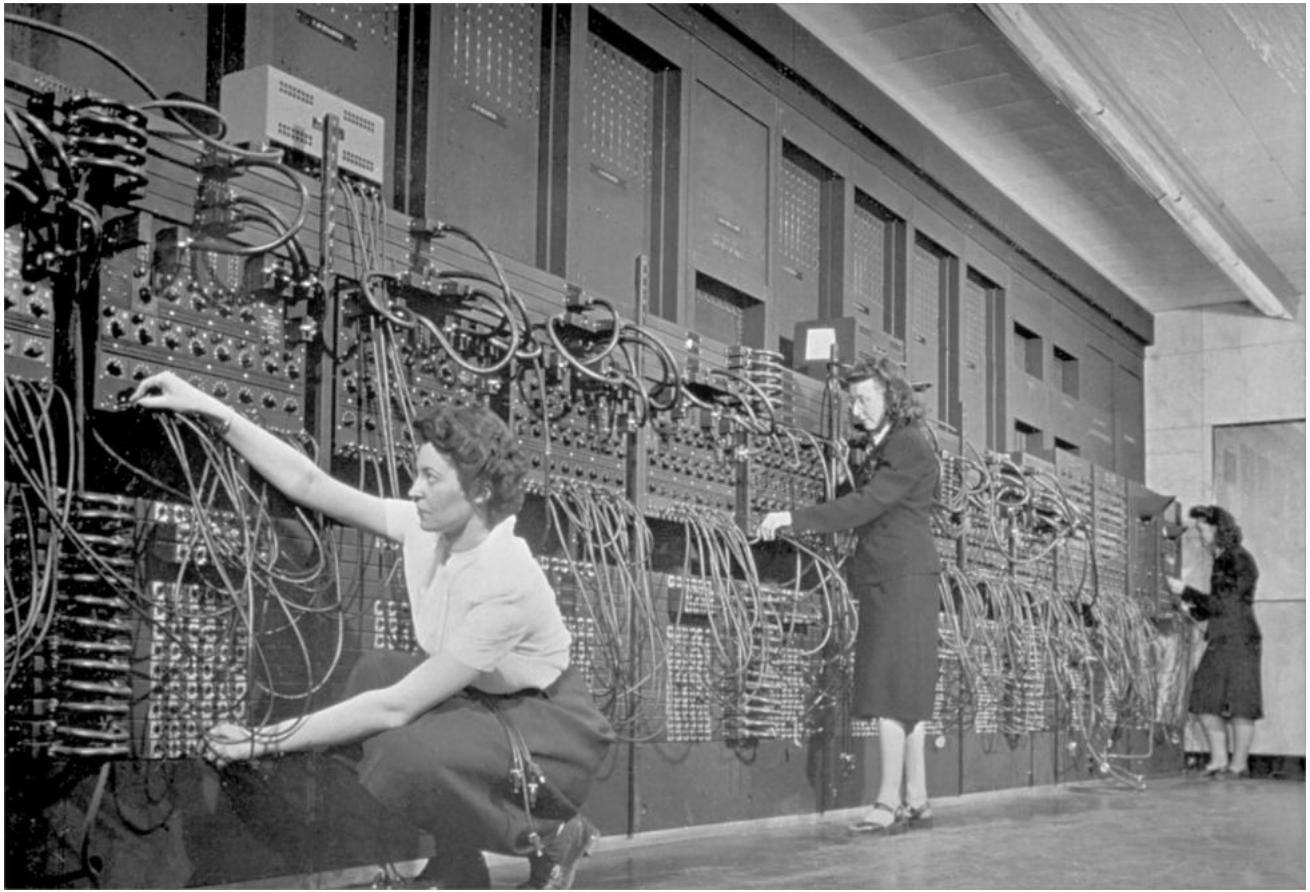


Pivotal

1946: ENIAC，第一台通用电子计算机

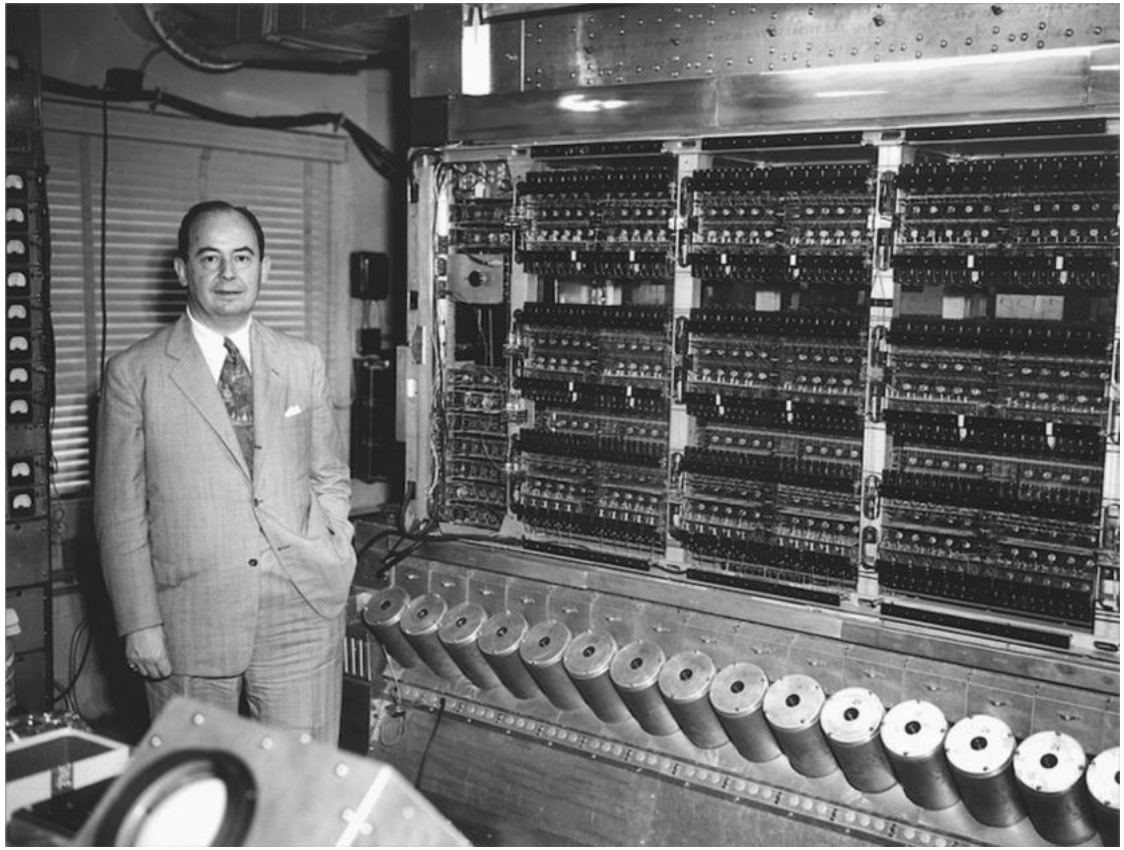


1946: ENIAC，编织程序(Wired Program)



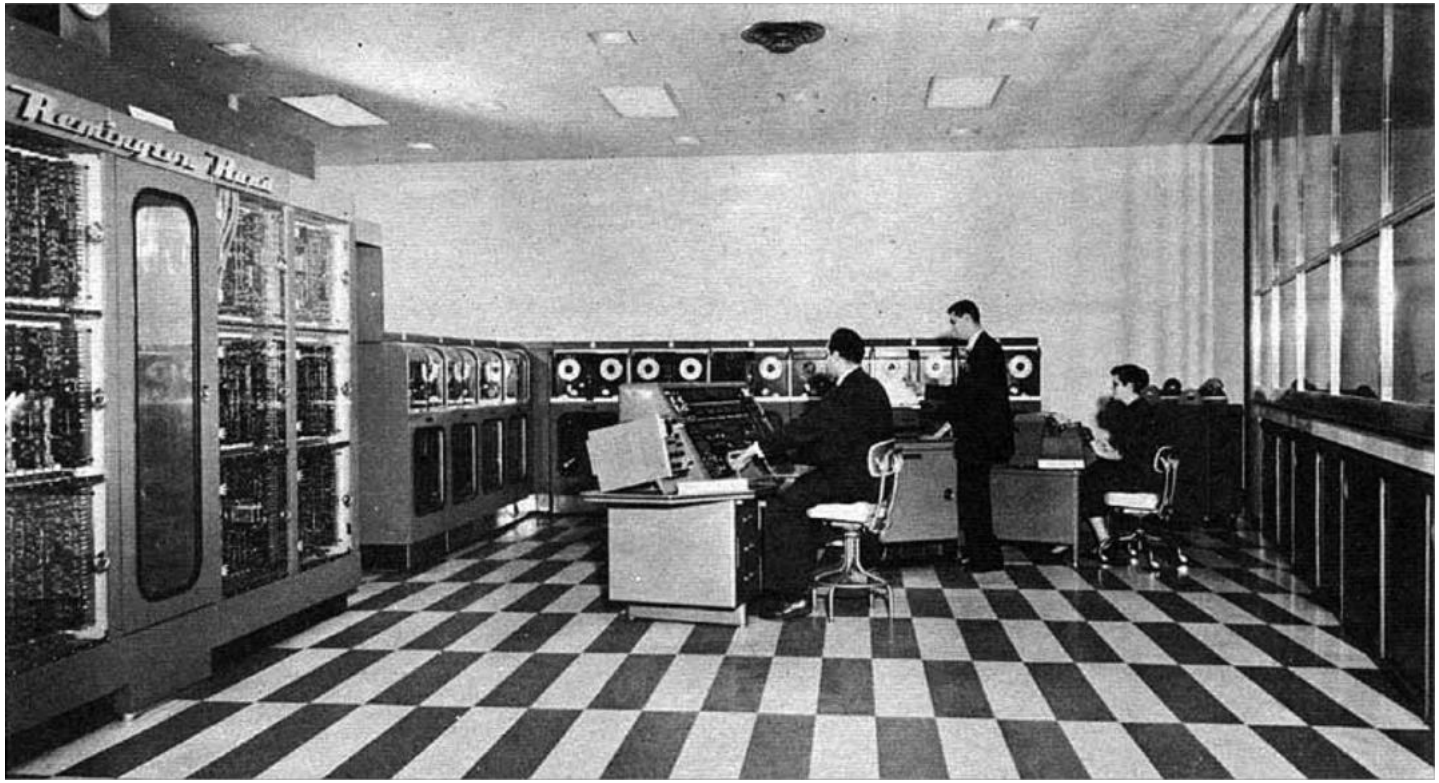
Pivotal

1949: EDVAC，存储程序原理



Pivotal

1951：UNIVAC，第一台商用计算机



Pivotal

操作系统

- 独占
- 批处理（监控程序）
- 分时共享
- 各种各样的OS: 通用为 IBM 704 开发 GM-NAA IO （1956）
- IBM 360 (1964)

1950s-1960s: 文件系统

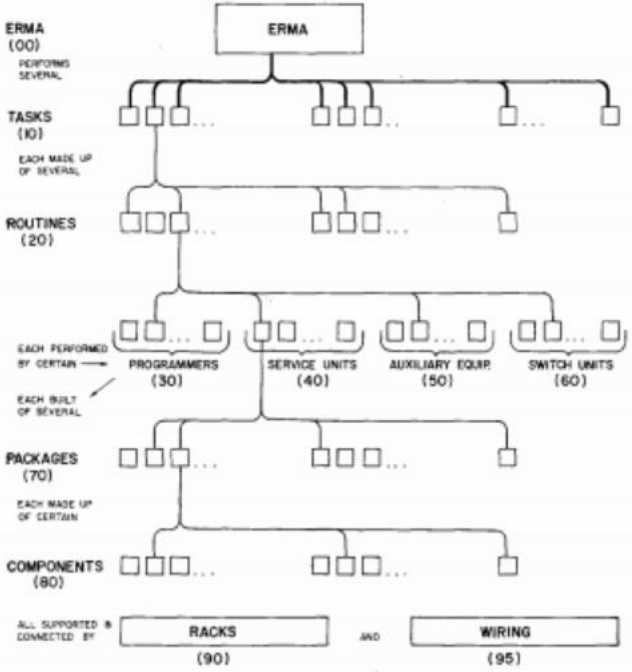


Fig. 2. The ERMA Mark I hierarchical structure

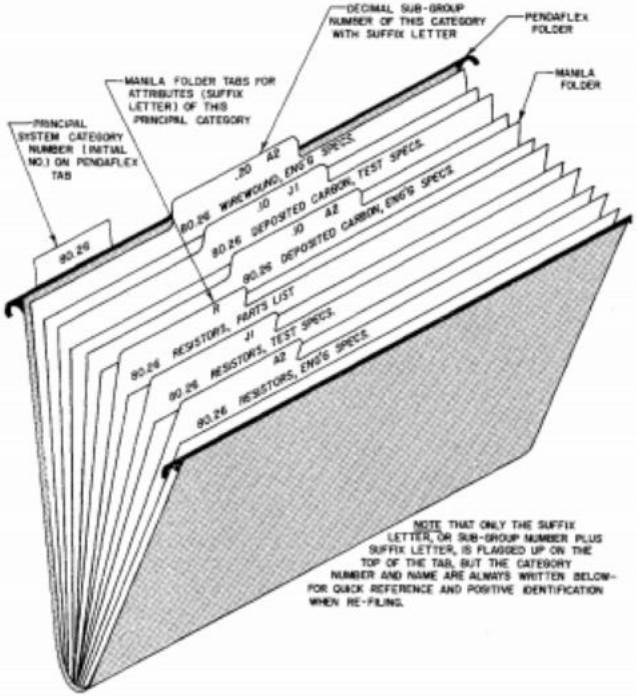


Fig. 3. Records storage and arrangement

数据库
实现第二次整合

Pivotal

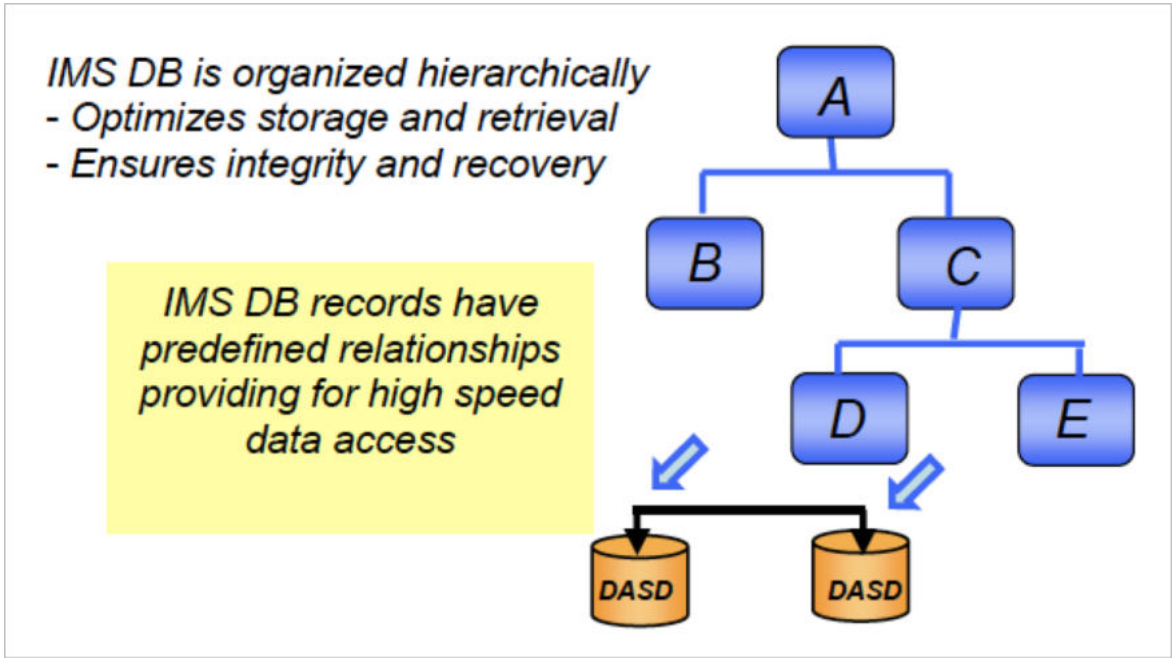
驱动需求

- 数据共享和整合
- 数据独立性
- 数据保护

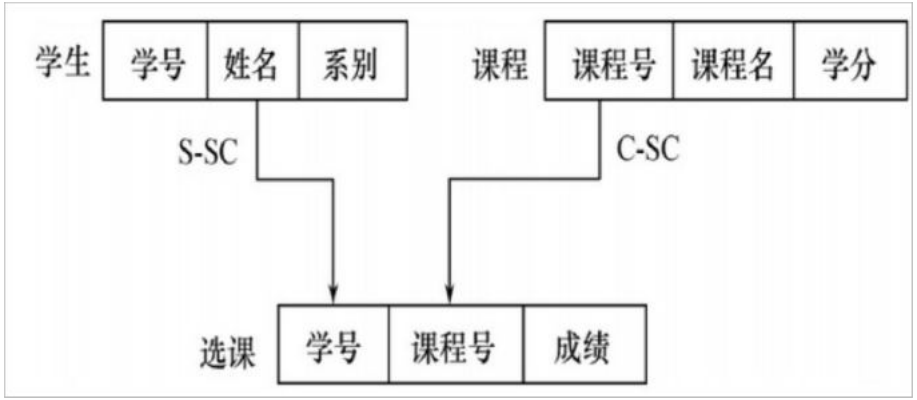
驱动需求

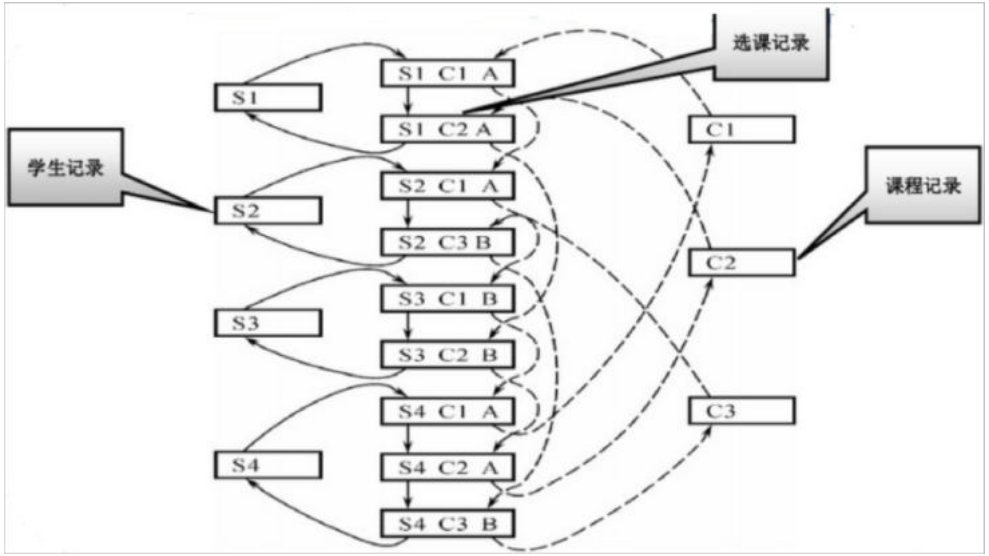
- 数据整合 → 数据模型（如何表示实体关系）
- 数据独立性 → 数据子语言
- 数据保护 → 日志、事务、权限

1968: IMS 层次数据库



网状数据库


学生
学号
姓名
系别
课程
课程号
课程名
学分
S-SC
C-SC
选课
学号
课程号
成绩


选课记录
学生记录
课程记录
S1
S2
S3
S4
C1
C2
C3
S1 C1 A
S1 C2 A
S2 C1 A
S2 C3 B
S3 C1 B
S3 C2 B
S4 C1 A
S4 C2 A
S4 C3 B

关系数据库

PARTS

PARTNO	NAME
P107	Bolt
P113	Nut
P125	Screw
P132	Gear

SUPPLIERS

SUPPNO	NAME
S51	Acme
S57	Ajax
S63	Amco

PRICES

PARTNO	SUPPNO	PRICE
P107	S51	.59
P107	S57	.65
P113	S51	.25
P113	S63	.21
P125	S63	.15
P132	S57	5.25
P132	S63	10.00

Fig. 1(b). A Relational Database.

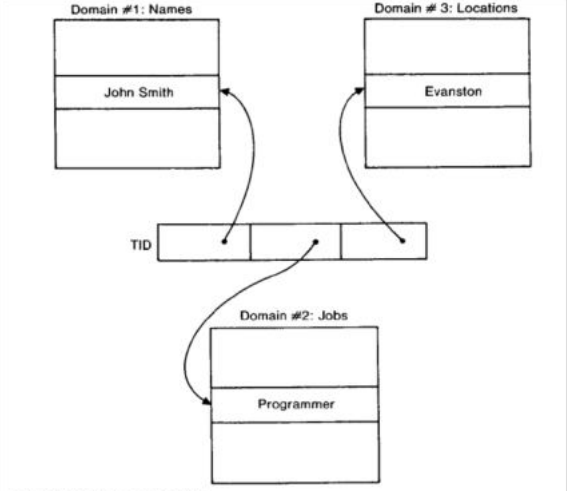


Fig. 2. XRM Storage Structure.

- System R
- Ingres
- Oracle
- IBM DB2：终结了2场竞争

对象数据库和关系对象数据库

- XML数据库曾是最火热的研究课题

- “站在前人的脚趾上，而不是站在前人的肩膀上” – Michael Stonebraker

初级阶段发展非常困难

60年代，数据库技术初现，欲取代文件系统：

- 慢
- 不稳定
- 人才匮乏
- 硬件能力

70年代初，关系模型当初也经历了很大的争论：

- 需要更多的系统资源，慢
- 人才匮乏
- 对优化技术的质疑

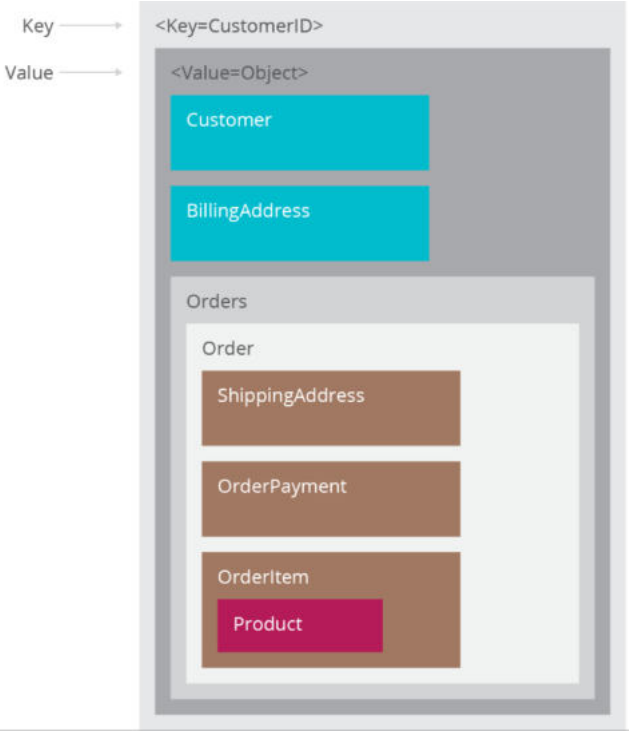
NoSQL 数据库

Pivotal

2005 - 今：225 个 NoSQL 产品

- 发展背景： 互联网 → 大数据 → 水平扩展
- 中间件： 底层大多使用关系数据库，ebay 的Oracle 集群，Google 的MySQL 集群
- MapReduce（2004）、BigTable（2006）、 Dynamo （2007）、Hadoop（2006）、 Cassandra（2008）、CouchDB（2005）、MongoDB（2007）、Hypertable（2008）、 Redis（2009）、ElasticSearch（2010）、Cassandra（2008）、MongoDB（2009）等。

2005-2015：K/V 数据模型



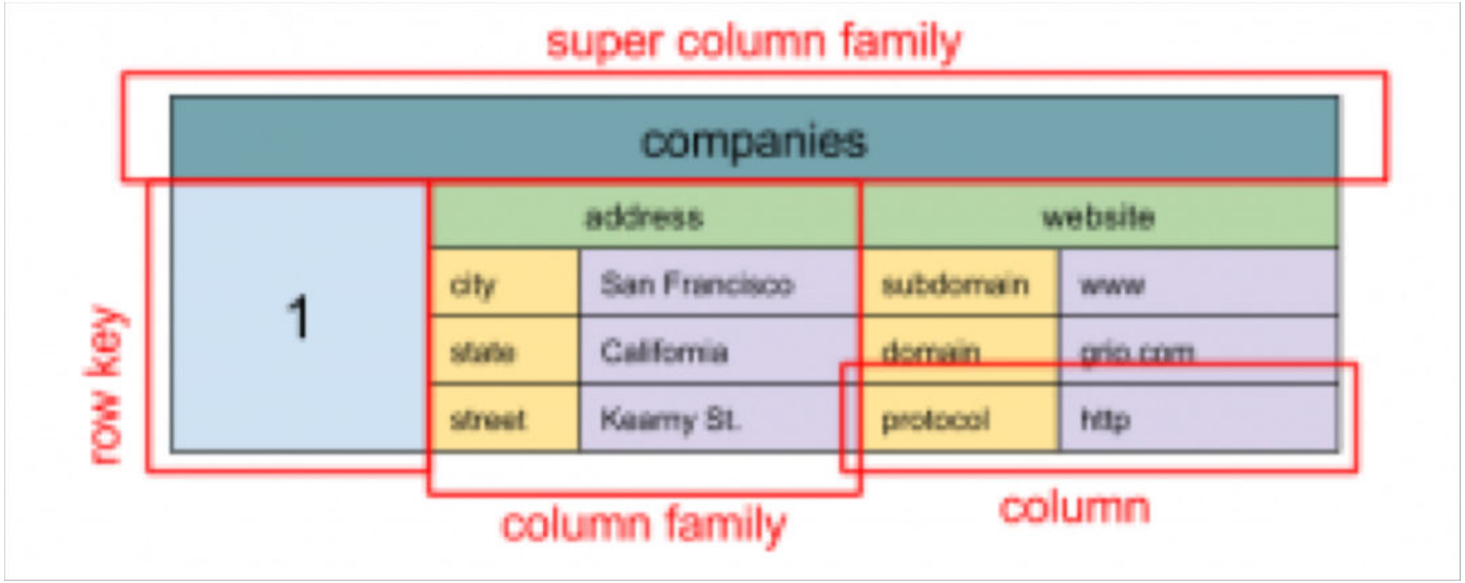
2005-2015：文档数据模型

<Key=CustomerID>

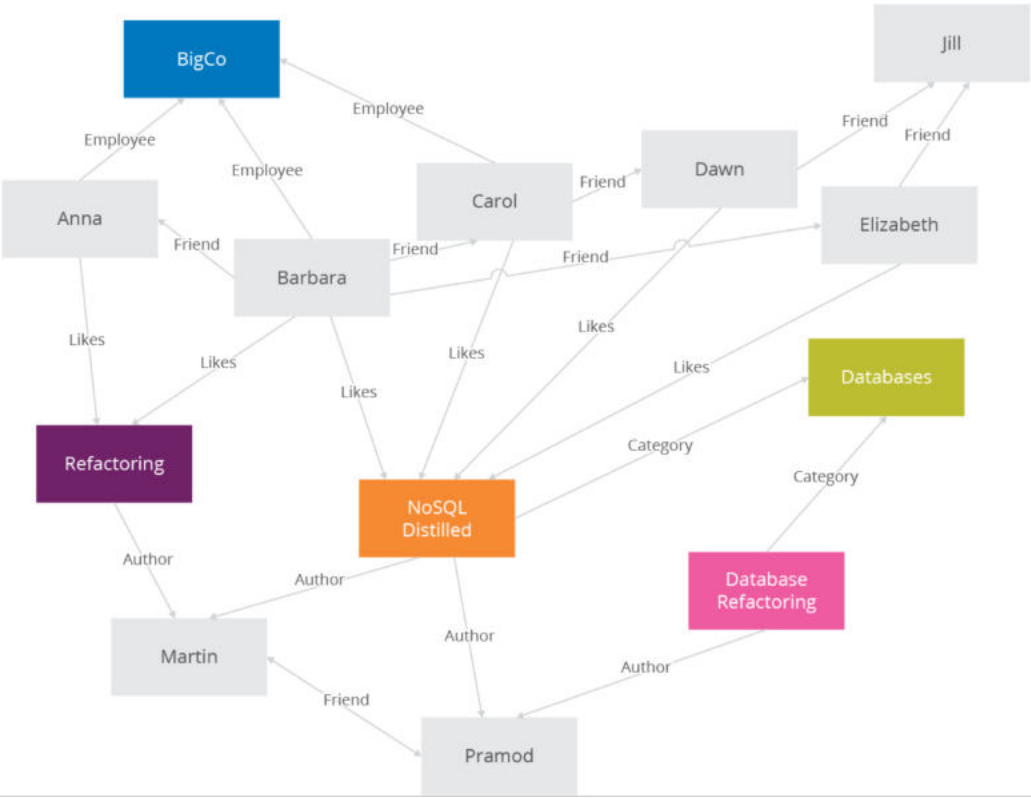
```
{
   "customerid": "fc986e48ca6"
   "customer":
   {
   "firstname": "Pramod",
   "lastname": "Sadalage",
   "company": "ThoughtWorks",
   "likes": [ "Biking","Photography" ]
   }
   "billingaddress":
   { "state": "AK",
      "city": "DILLINGHAM",
      "type": "R"
    }
}
```

Key

2005-2015：列数据库



2005-2015: Graph 数据模型



SQL 数据库回归？

Pivotal

2008：MapReduce 是技术大倒退

作为一种数据处理模式，MapReduce 是一种大倒退。

- 模式（Schema）很有价值
- 从应用程序中抽离出模式利于共享
- 高级访问语言很有价值

NoSQL 的局限性

- 不支持SQL，开发人员自己实现复杂的代码，进行聚集分析等。
- 不支持ACID和事务，实现大量代码处理数据不一致
- 不支持关联，只得使用宽表，引起数据冗余，维护代价高
- 使用低级查询语言，数据独立性差，灵活性差，维护代价高。
- 缺少标准接口，学习代价高，应用使用代价高，需要大量胶水代码
- 缺少生态，从数据迁移、ETL、报表、BI、可视化都要从头开发
- 多种 NoSQL 产品的引入，数据整合代价高
- 人才缺乏，企业积累的大量SQL人才和资产浪费

1980s - 2018: SQL 从未离开

- NoSQL的主要推动力是大数据引起的集群化需求和线性扩展能力以获得更高的性能和可用性。
- 分布式集群式SQL数据库的研究可以追溯到20世纪80年代中期。包括 Gamma、Teradata、Bubba和Tandem 等。
- 目前主流的分布式SQL和NoSQL数据库都采用无共享架构。

2010 前后

- OLAP分布式SQL：Netezza（2003）、Greenplum（2005）、Vertica（2005）、HANA（2008）

- OLTP分布式SQL：H-Store （2007）、VoltDB （2010）、ClustrixDB(2006)、CockroachDB (2015) → NewSQL, New Vendor

集成数据处理平台
第三次整合

Pivotal

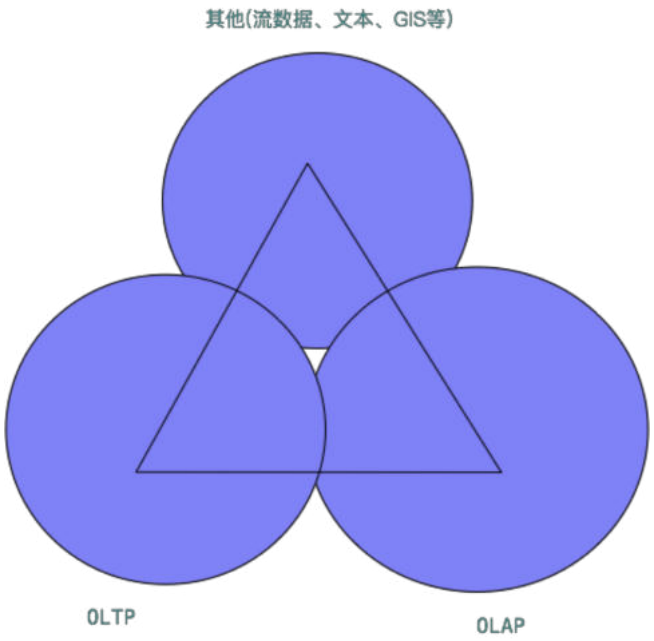
处理不同数据类型

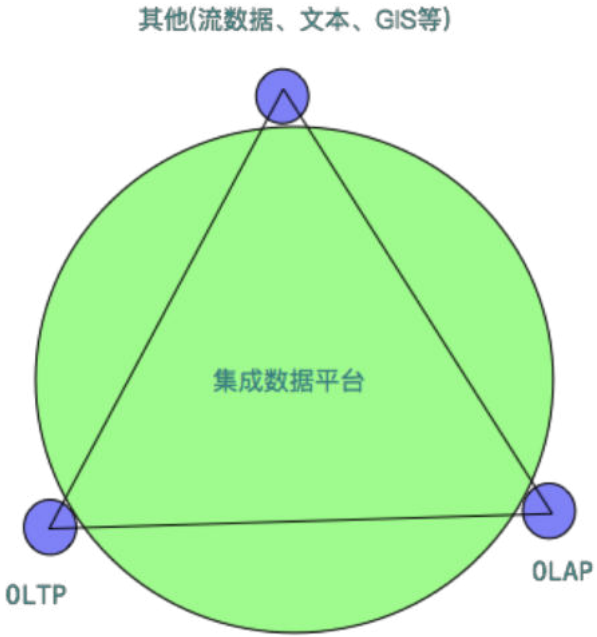
- 结构化
- 文本数据
- 地理信息数据
- XML
- JSON
- 图数据
- 音视频

处理不同业务场景

- HTAP（OLTP + OLAP）
- 流处理
- 数据分析
- 其他等

分还是∑





需求和技术是一对矛盾

- 该矛盾缓和时，数据处理领域将更趋向于整合；
- 该矛盾尖锐时，数据处理领域将趋于分散

就软硬件技术发展现状来看，未来整合的趋势更为明显。

处理整合带来的新问题

- 数据和处理进一步整合：分布式和数据库两个领域的碰撞
- 数据保护：高可用、安全（多租户）
- 独立性：SQL

Greenplum：从MPP数据库到大数据处理基础平台

Pivotal

Morgan Stanley Ford Go Further IRS Bloomberg 中信银行 CHINA CITIC BANK sas INFORMATICA The Data Integration Company COGNOS

Pivotal Labs THE LEAN STARTUP Lean startup cycle USER CENTERED DESIGN Agile

GREENPLUM DATABASE PostgreSQL ORCA MADlib APACHE GEODE

BI TEXT GIS Regression Classification Clustering Graph Image

SQL 1992/1998/2003/2008 完善的 SQL 支持 ODBC Java JDBC OLEDB

SQL python R THE C PROGRAMMING LANGUAGE Extensions UDT/UDA PXF External Table

数据分布 并行执行 优化器 线性扩展 大规模并行处理 (MPP) 多态存储 资源管理 高可用 高速加载

结构化数据 JSON XML 半结构化 文本数据 图数据 GIS数据 图像数据

hadoop HIVE APACHE HBASE APACHE Spark APACHE kafka Pivotal GemFire PostgreSQL Pivotal Greenplum Java JDBC

CF Pivotal CF amazon web services Google Cloud Platform Microsoft Azure 阿里云 aliyun.com 腾讯云 vmware vSphere openstack.

企业级稳定性、成熟生态系统

敏捷方法学：快速迭代、持续发布、质量内建

开放源代码、其持续大力投入

集成数据平台：BI、文本、GIS、图、图像、流式

完善的标准支持：SQL、JDBC、ODBC

强大的灵活性、可扩展

大规模并行处理 (MPP): 强大内核

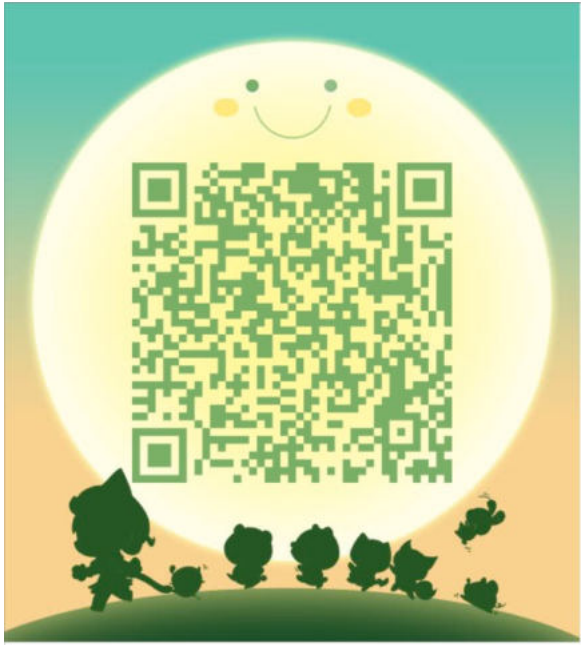
各种数据格式：结构化、半结构化、非结构化

各种数据源：Hadoop、S3、数据库、文件

一次打包，到处运行：裸机、私有云、公有云

Greenplum 所处的时代

- 巨大机会：仍然处在百家争鸣、封疆裂土的战国时代，市场已经确认
- 巨大挑战：竞争激烈、玩家众多、新型玩法、新兴硬件技术
- Pivotal 的战略和愿景



GreenplumDB

扫一扫二维码，加入群聊。