

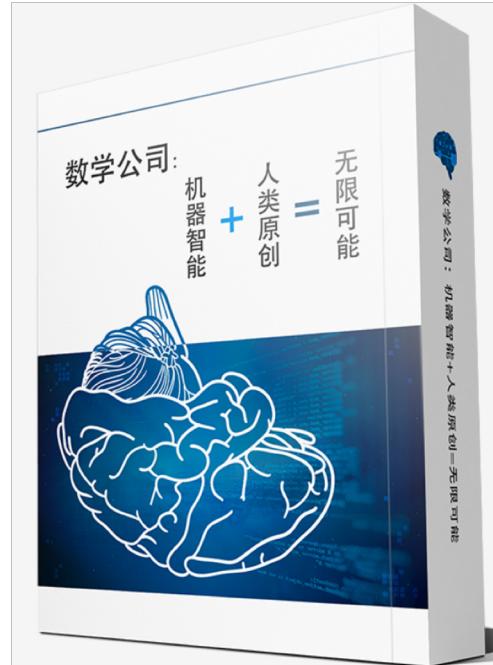
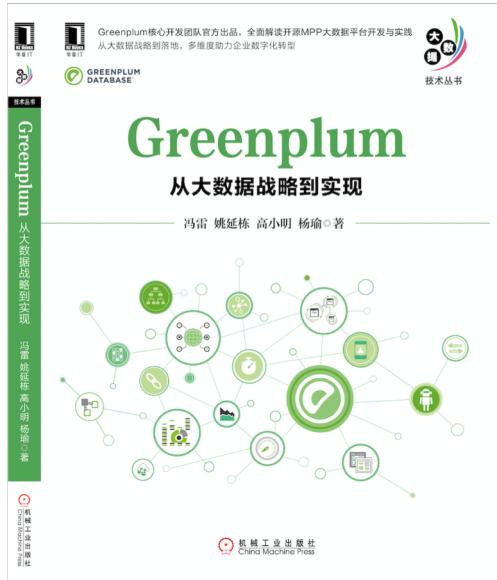
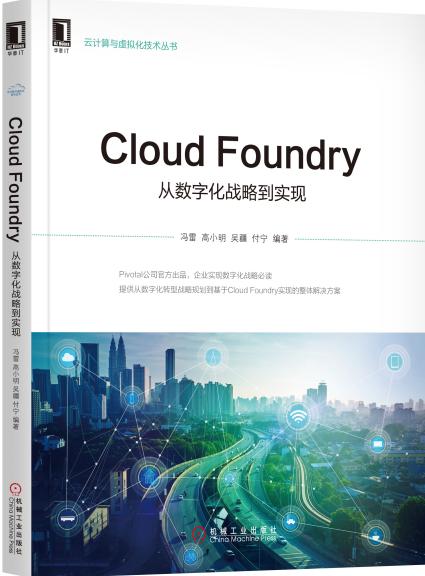
忘掉 Hadoop 大数据 ≈ 分布式数据库

姚延栋

yyao@pivotal.io

关于 Pivotal 商业逻辑和产品

数字化转型三部曲



A photograph of a professional environment. In the foreground, a man with a beard and a light-colored shirt stands on the right, looking towards the left. In the center, another man is writing on a whiteboard with a marker. Several other people are seated or standing in the background, engaged in conversation or observation. The room has a modern design with exposed ceiling beams and large windows.

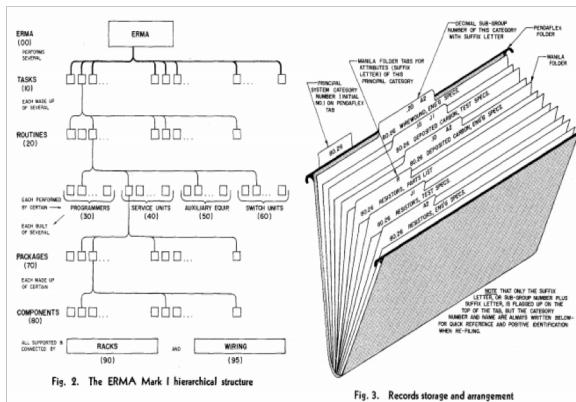
从无到有演进看大数据发展

“史前时代”：文件系统

1950 1960 1970 1980 1990 2000 2010 2019

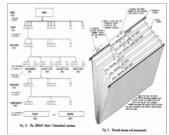
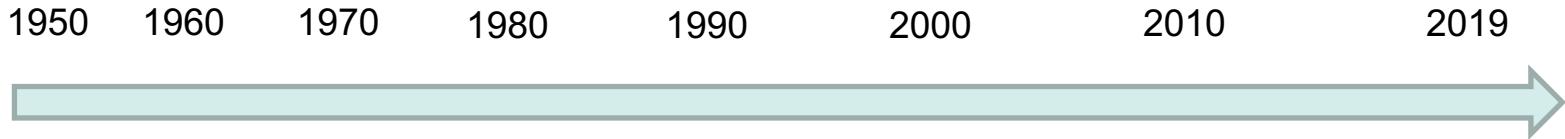


文件系统 Flat 文件格式

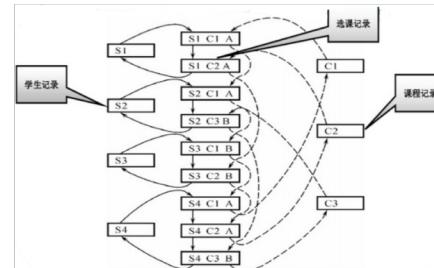
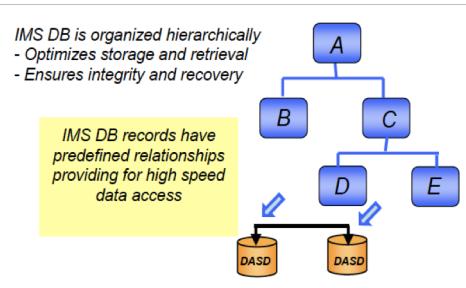


Pivotal

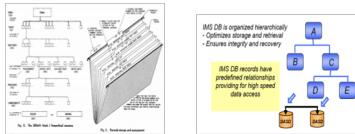
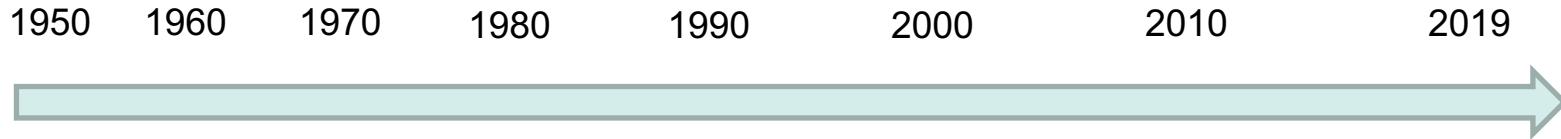
六七十年代：层状数据库和网状数据库



IMS 网状数据库



八九十年代至今：关系数据库

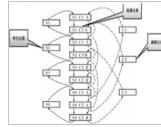
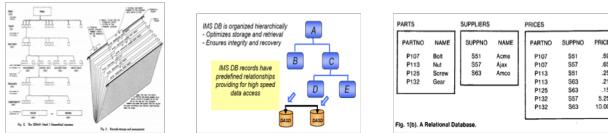
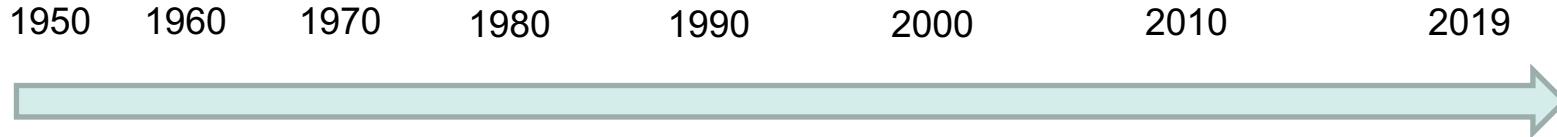


关系数据库

PARTS		SUPPLIERS		PRICES		
PARTNO	NAME	SUPPNO	NAME	PARTNO	SUPPNO	PRICE
P107	Bolt	S51	Acme	P107	S51	.59
P113	Nut	S57	Ajax	P107	S57	.65
P125	Screw	S63	Amco	P113	S51	.25
P132	Gear			P113	S63	.21
				P125	S63	.15
				P132	S57	5.25
				P132	S63	10.00

Fig. 1(b). A Relational Database.

八十年代：对象数据库



对象数据库

Object-Oriented Model

Object 1: Maintenance Report Object 1 Instance

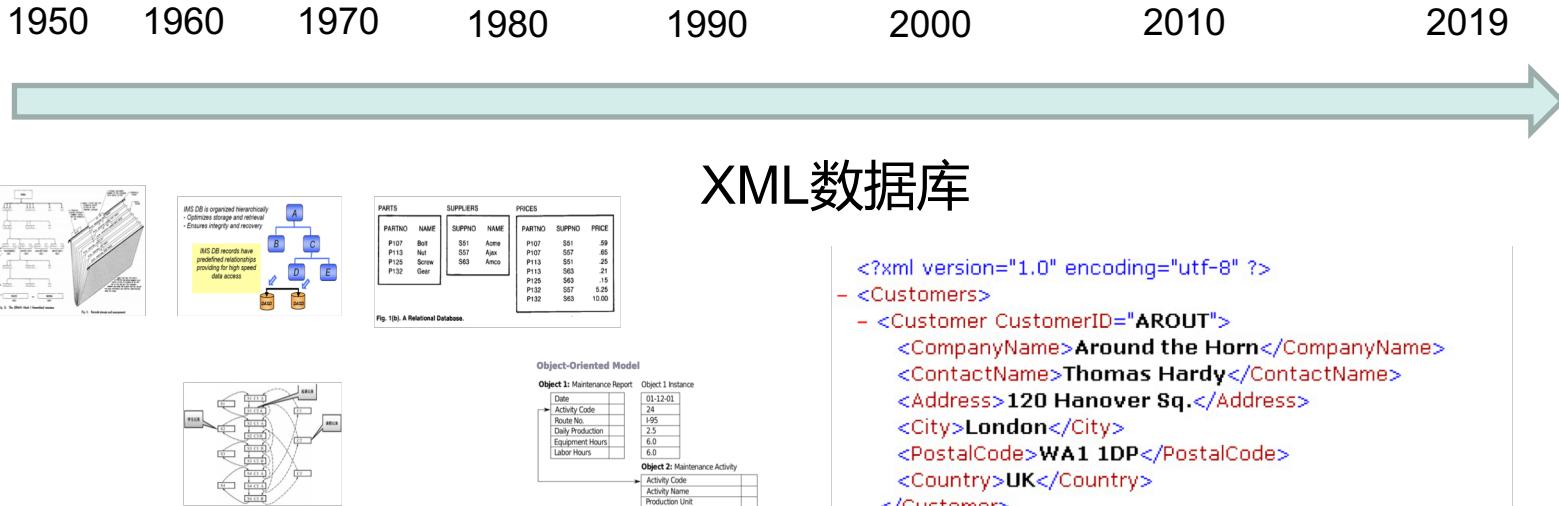
Date	
Activity Code	
Route No.	
Daily Production	
Equipment Hours	
Labor Hours	

01-12-01
24
I-95
2.5
6.0
6.0

Object 2: Maintenance Activity

Activity Code	
Activity Name	
Production Unit	
Average Daily Production Rate	

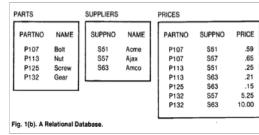
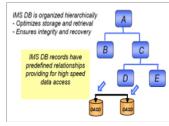
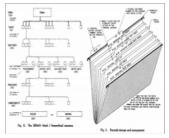
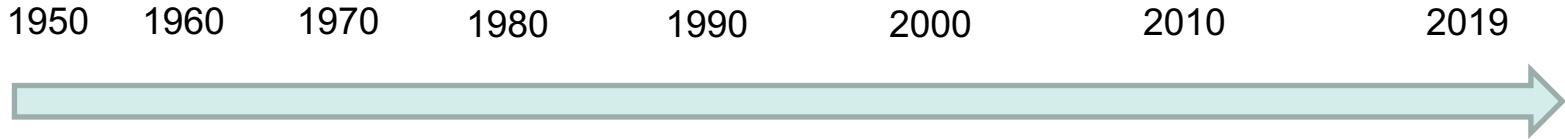
九十年代千年：XML 数据库



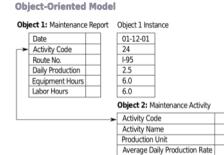
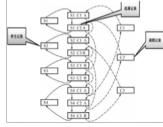
XML数据库

```
<?xml version="1.0" encoding="utf-8" ?>
- <Customers>
  - <Customer CustomerID="AROUT">
    <CompanyName>Around the Horn</CompanyName>
    <ContactName>Thomas Hardy</ContactName>
    <Address>120 Hanover Sq.</Address>
    <City>London</City>
    <PostalCode>WA1 1DP</PostalCode>
    <Country>UK</Country>
  </Customer>
  - <Customer CustomerID="CHOPS">
    <CompanyName>Chop-suey Chinese</CompanyName>
    <ContactName>Yang Wang</ContactName>
    <Address>Hauptstr. 29</Address>
    <City>Bern</City>
```

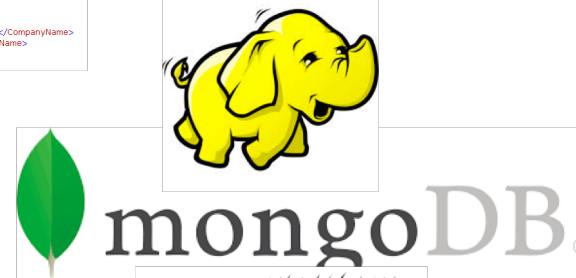
2005-2017 : NoSQL



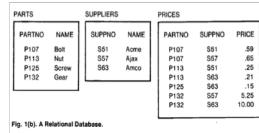
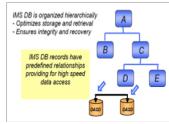
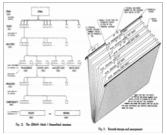
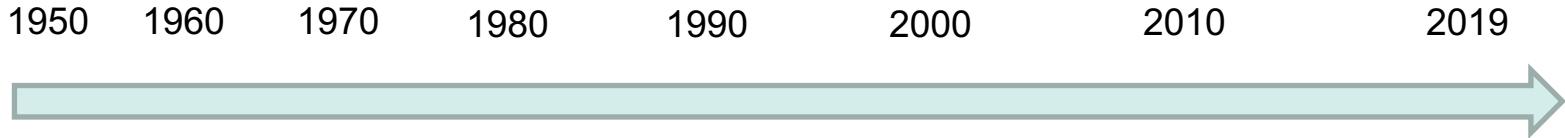
<?xml version="1.0" encoding="utf-8"?>
- <Customers>
- <Customer CustomerId="AROUT">
- <CompanyName>Around the Horn</CompanyName>
- <ContactName>Thomas Hardy</ContactName>
- <Address>120 Hanover Sq.</Address>
- <City>London</City>
- <PostalCode>WA1 1DP</PostalCode>
- <Country>UK</Country>
</Customer>
- <Customer CustomerId="CHOPS">
- <CompanyName>Chop-suey Chinese</CompanyName>
- <ContactName>Yang Wang</ContactName>
- <Address>Hauptstr. 29</Address>
- <City>Bern</City>



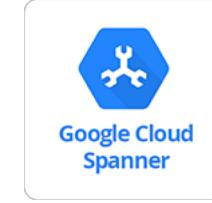
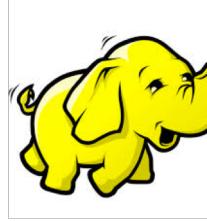
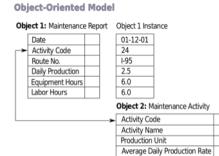
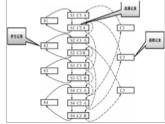
NoSQL



SQL 回归 ?

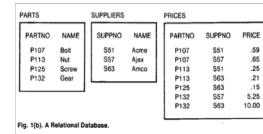
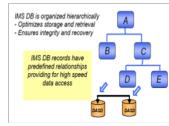
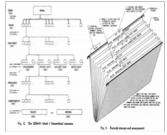
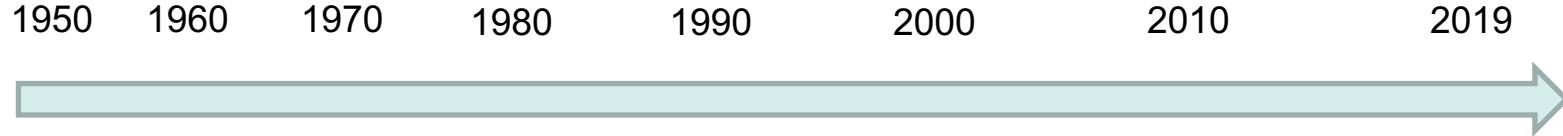


```
<?xml version="1.0" encoding="utf-8"?>
- <Customers>
  - <Customer CustomerId="AROUT">
    <CompanyName>Around the Horn</CompanyName>
    <ContactName>Thomas Hardy</ContactName>
    <Address>120 Hanover Sq.</Address>
    <City>London</City>
    <PostalCode>WA1 1DP</PostalCode>
    <Country>UK</Country>
  </Customer>
  - <Customer CustomerId="CHOPC">
    <CompanyName>Chop-suey Chinese</CompanyName>
    <ContactName>Wong Wang</ContactName>
    <Address>Hauptstr. 29</Address>
    <City>Bern</City>
```

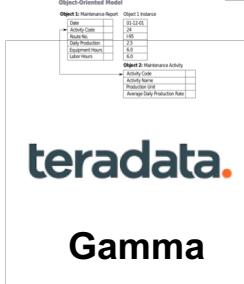
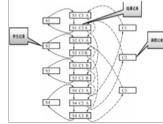


SQL回归 ?

SQL 从未离开



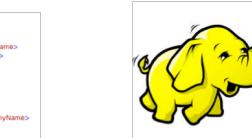
```
<xml version="1.0" encoding="utf-8" >
- <Customers>
  - <Customer CustomerId="A40101">
    <CompanyName>Around the Horn</CompanyName>
    <Address>12悠然街 8号</Address>
    <City>London</City>
    <PostalCode>WA1 1DP</PostalCode>
    <Country>UK</Country>
  </Customer>
  <Customer CustomerId="C10001">
    <CompanyName>Chop-sky Chinese</CompanyName>
    <ContactName>Yang Wang</ContactName>
    <Address>Haupstr. 29</Address>
    <City>Bern</City>
  </Customer>
```



Gamma

Tandem

Pivotal



Pivotal
Greenplum®

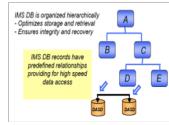
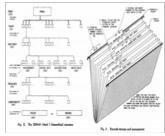
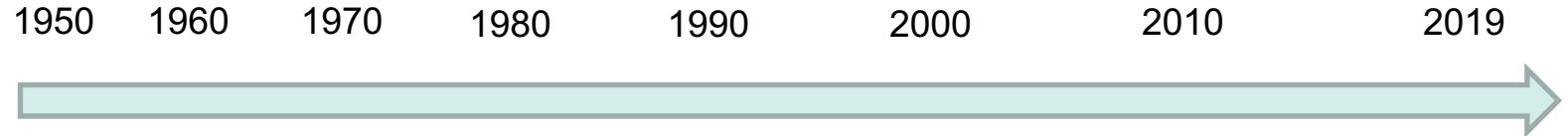
SAP HANA

VERTICA

VOLTDB

Cockroach DB

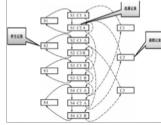
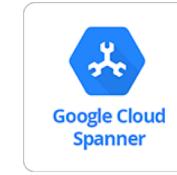
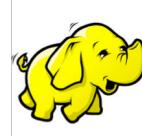
MEMSQL



PARTNO	NAME	SUPPNO	NAME	PARTNO	SUPPNO	PRICE
P107	Bolt	551	Arcia	P107	551	.59
P113	Nut	557	Ajax	P107	557	.65
P125	Screw	563	Amco	P113	551	.25
P132	Clamp			P125	563	.15
P132				P132	567	5.25
P132				P132	563	10.20

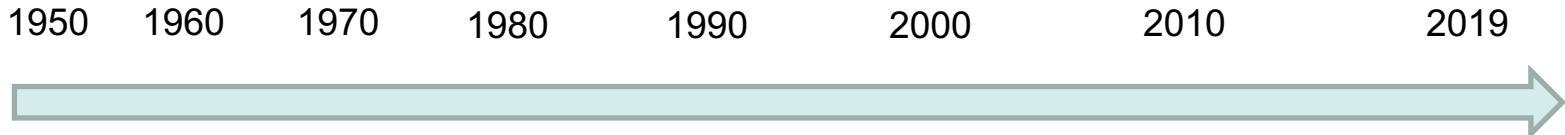
Fig. 1(b). A Relational Database.

```
<xml version="1.0" encoding="utf-8" >
- <Customer>
  - <Customer> Customer="AROUT"-
    <CompanyName>Around the Horn</CompanyName>
    <ContactName>Thomas Hardy</ContactName>
    <Address>123 Main St</Address>
    <City>London</City>
    <PostalCode>WA1 1DP</PostalCode>
    <Country>GBR</Country>
- <Customer>
  <Customer> Customer="CHOPP"-
    <CompanyName>Chop-sky Chinese</CompanyName>
    <ContactName>Yang Wang</ContactName>
    <Address>Haupstr. 29</Address>
    <City>Bern</City>
```



分 布 式 数 据 库

过去60年大约每十年一次大争论



文件系统 vs. 数据库

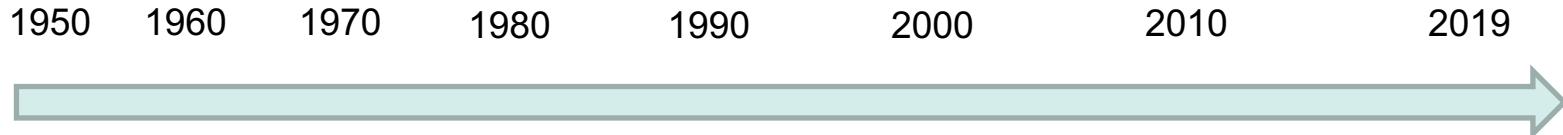
CODSAL vs. 关系数据库

对象数据库 vs. 关系数据库

XML 数据库 vs. 关系数据库

NoSQL vs. 分布式关系数据库

数据库、关系数据库、分布式关系数据库 胜出



文件系统 vs. 数据库

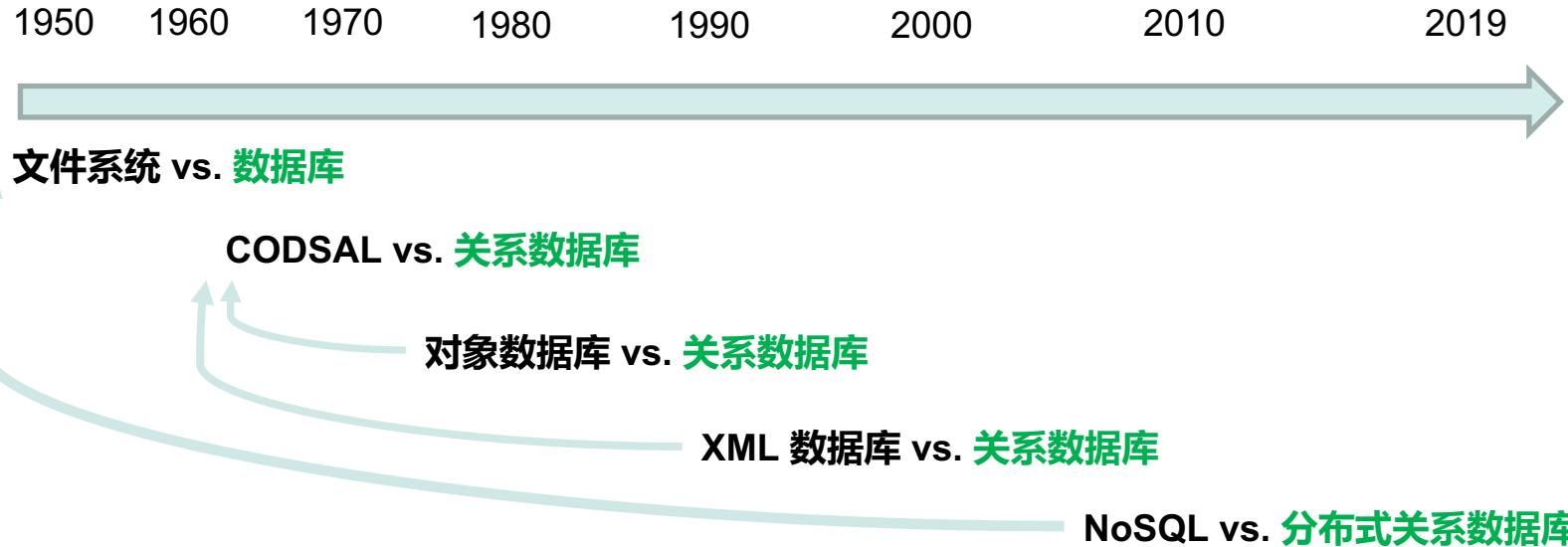
CODSAL vs. 关系数据库

对象数据库 vs. 关系数据库

XML 数据库 vs. 关系数据库

NoSQL vs. 分布式关系数据库

历史在不断的重复



数据库发展原动力

- 数据共享和整合
- 数据独立性
- 数据保护

数据库原动力

- 数据整合 → 数据模型
- 数据独立性 → 数据子语言
- 数据保护 → 日志、事务、权限

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
- 实现糟糕，不支持 Access Method
- 高级访问语言很有价值

Google 内部2011年放弃MR，2015年公开放弃

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

Hadoop 市场是SQL市场，是分析型数据市场

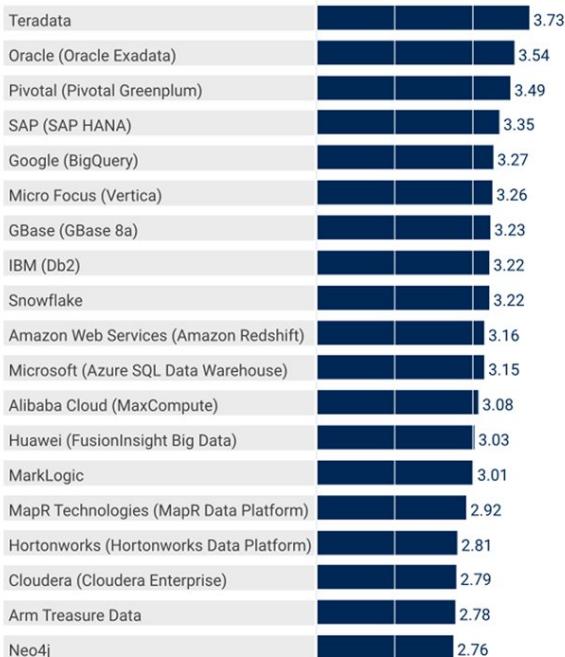
- Hadoop 含义的演进： HDFS/MR/Hive/Hbase
- Hadoop 发布在技术未成熟前已经过时（ Gartner 2017 ）
- 70% 的 Hadoop 部署未达成目标（整合困难，技能不足）
- Strata+Hadoop → Strata (2018 年)
- Cloudera : 75% 的 Hadoop 市场是 SQL 市场,
- Facebook: 95+% Hive
- Spark : SparkSQL 70%

Hadoop 市场是SQL市场，然而其在这个市场中无优势

- Greenplum 全球排名第三
- 前十中唯一的开源产品
- 实时分析领域排名并列第四

Figure 1. Vendors' Product Scores for Traditional Data Warehouse Use Case

Product or Service Scores for Traditional Data Warehouse



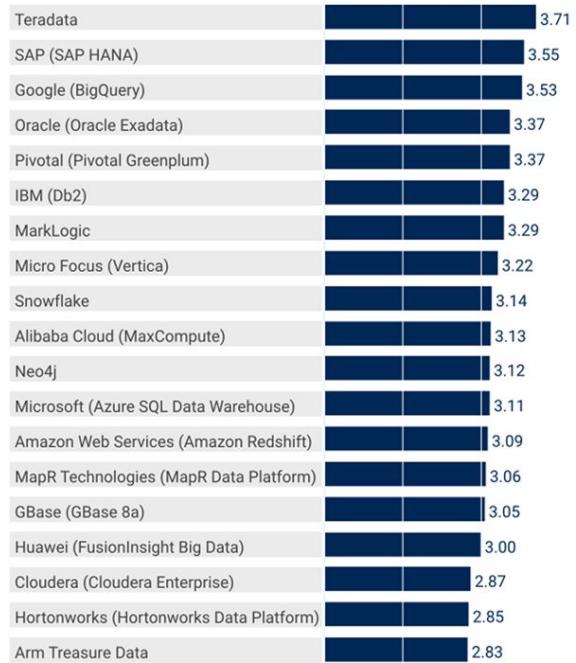
As of 21 January 2019

Source: Gartner (March 2019)

Pivotal

Figure 2. Vendors' Product Scores for Real-Time Data Warehouse Use Case

Product or Service Scores for Real-Time Data Warehouse



As of 21 January 2019

Source: Gartner (March 2019)

© Gartner, Inc.

大数据 ≈ 分布式数据库



Pivotal
Greenplum[®]



GreenplumDB

扫一扫二维码，加入群聊。