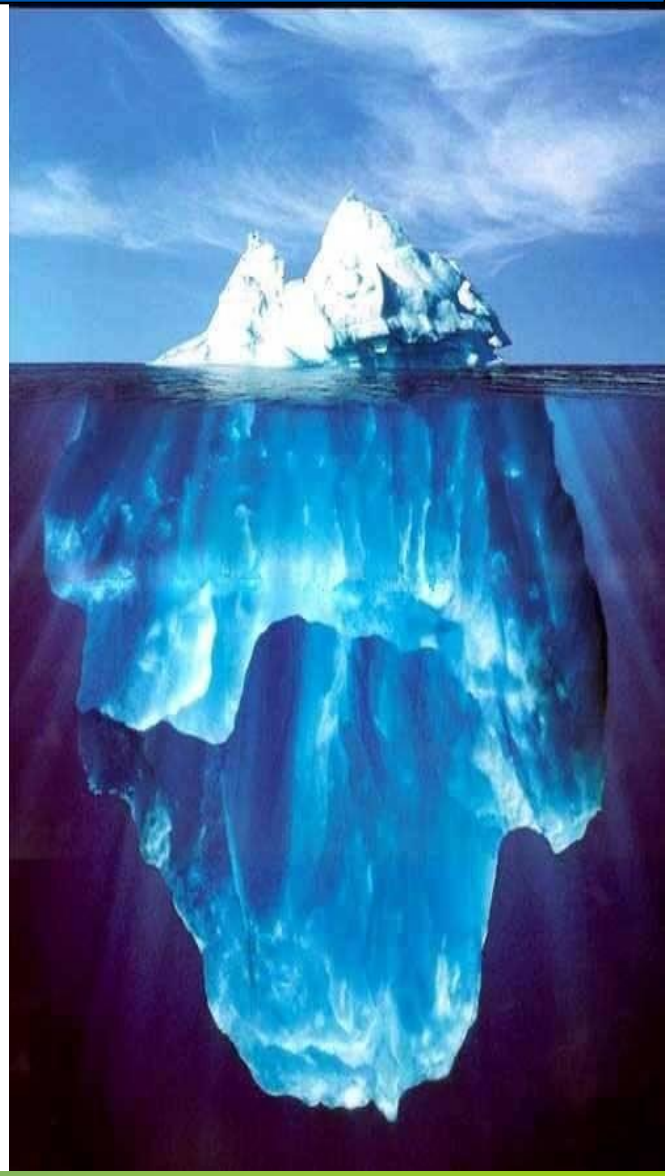
第2章 大数据关键技术与挑战
（2013年新版）

E-mail: ziyulin@xmu.edu.cn

http://www.cs.xmu.edu.cn/linziyu

提纲

- 大数据处理的基本流程
- 大数据处理模型
- 大数据关键技术
- 大数据处理工具
- 大数据时代面临的新挑战

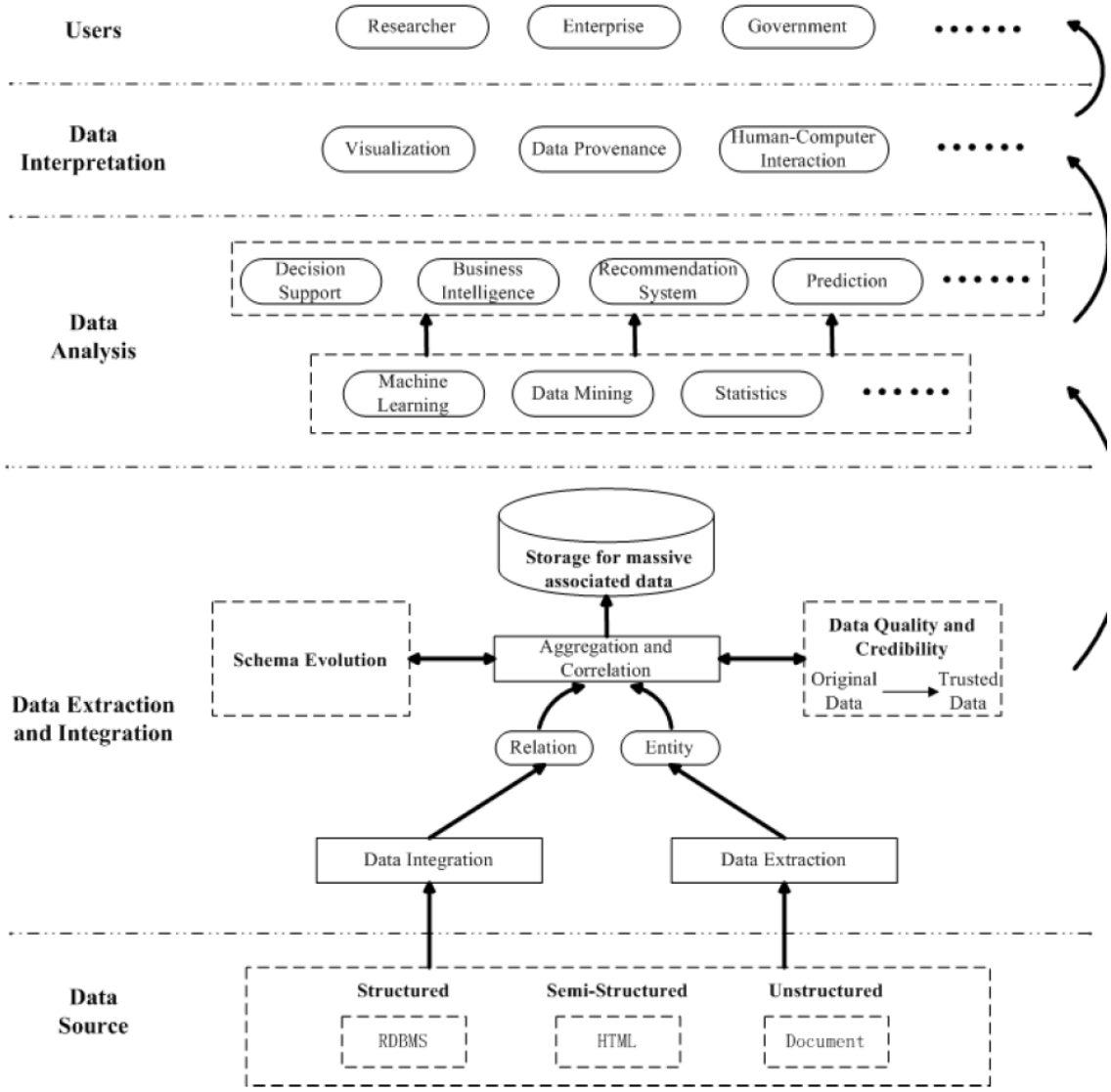


PPT

2013

http://dblab.xmu.edu.cn/node/423

大数据处理的基本流程


Users
Researcher
Enterprise
Government
Data Interpretation
Visualization
Data Provenance
Human-Computer Interaction
Data Analysis
Decision Support
Business Intelligence
Recommendation System
Prediction
Machine Learning
Data Mining
Statistics
Storage for massive associated data
Schema Evolution
Aggregation and Correlation
Data Quality and Credibility
Original Data
Trusted Data
Data Extraction and Integration
Relation
Entity
Data Integration
Data Extraction
Data Source
Structured
Semi-Structured
Unstructured
RDBMS
HTML
Document

数据抽取与集成

-
-
-
-

数据分析

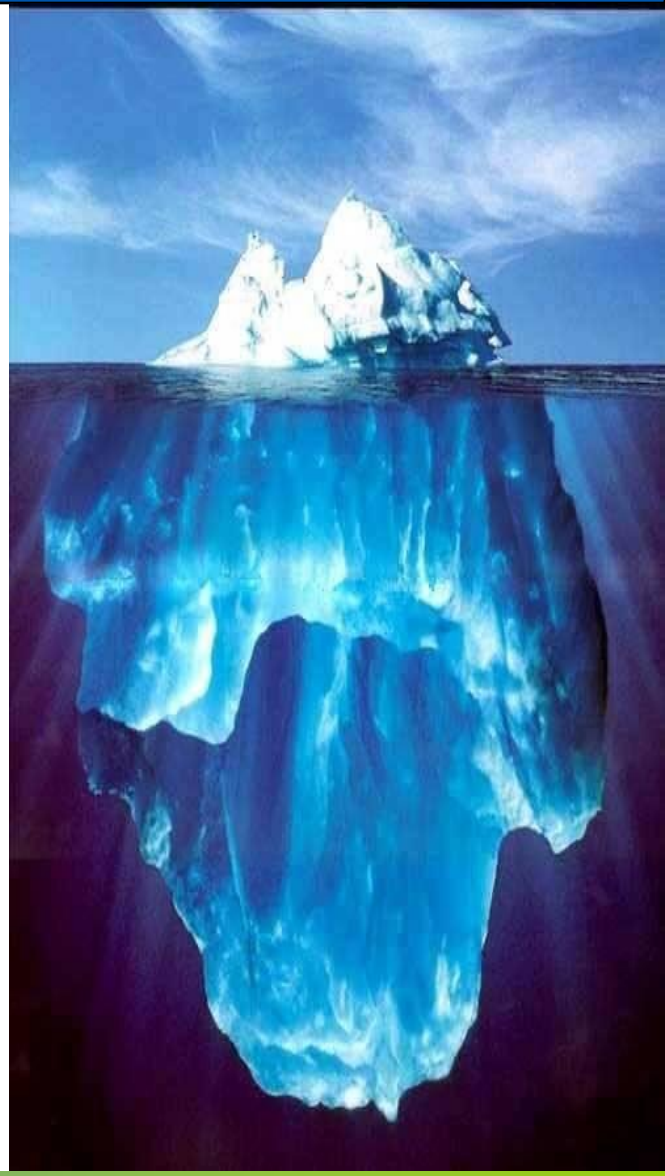
-
  -
  -
  -

数据解释

-
-
-
  -
  -

提纲

- 大数据处理的基本流程
- 大数据处理模型
- 大数据关键技术
- 大数据处理工具
- 大数据时代面临的新挑战



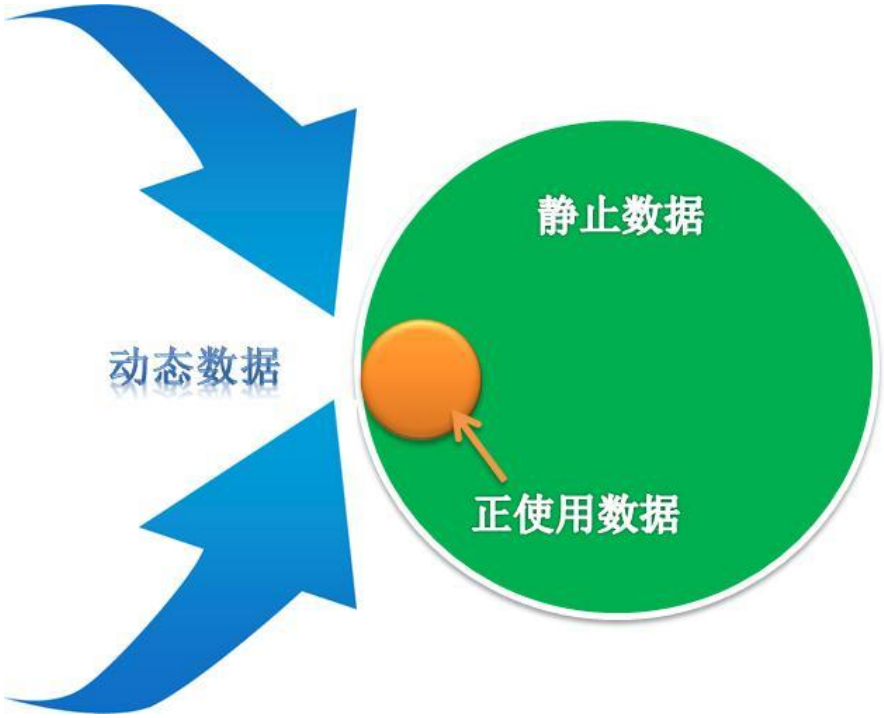
大数据之“快”从何说起

时间就是金钱

像其它商品一样，数据的价值会折旧

数据跟新闻和金融行情一样，具有时效性

大数据的三种状态



“ data at rest ”

“ data inuse ”

“ data in motion ”

大数据的“快”说的是两个层面

- “ ”

  burst

  CERN

  (Large Hadron Collider LHC)

  PB

  clickstream

  RFID GPS Twitter firehose

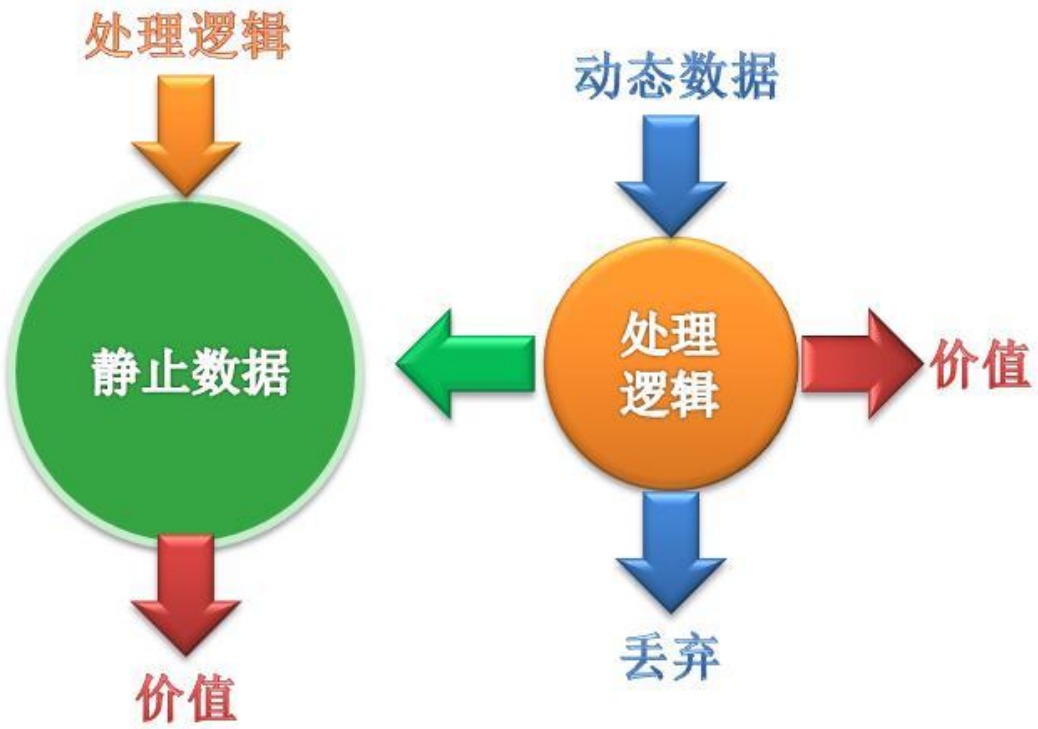
- “ ”

  “

  ” “ ”

  “ ” “ ”

批处理与流处理



“ ”

“ ”

“ ”

批处理与流处理的组合

- PB

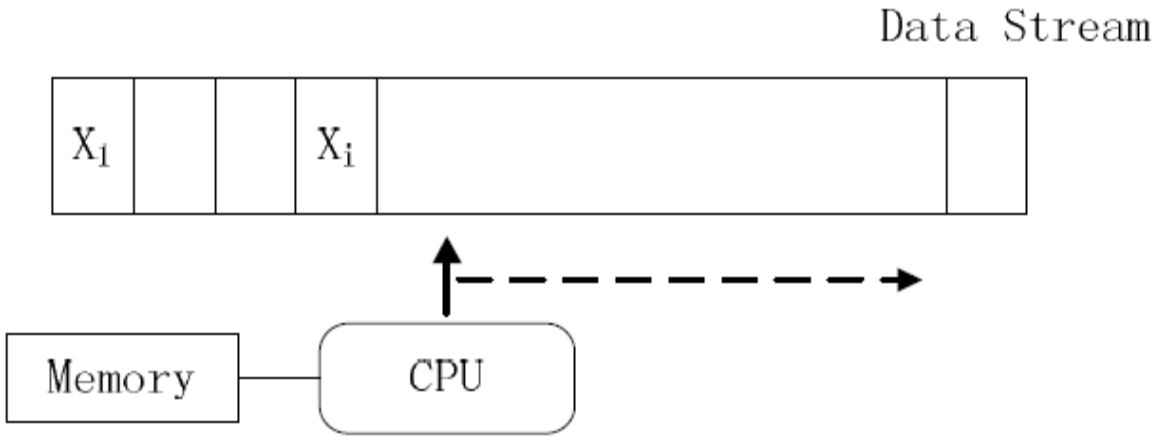
  25PB

-

如何实现“快”的数据处理

-
-
- --
- I/O
-

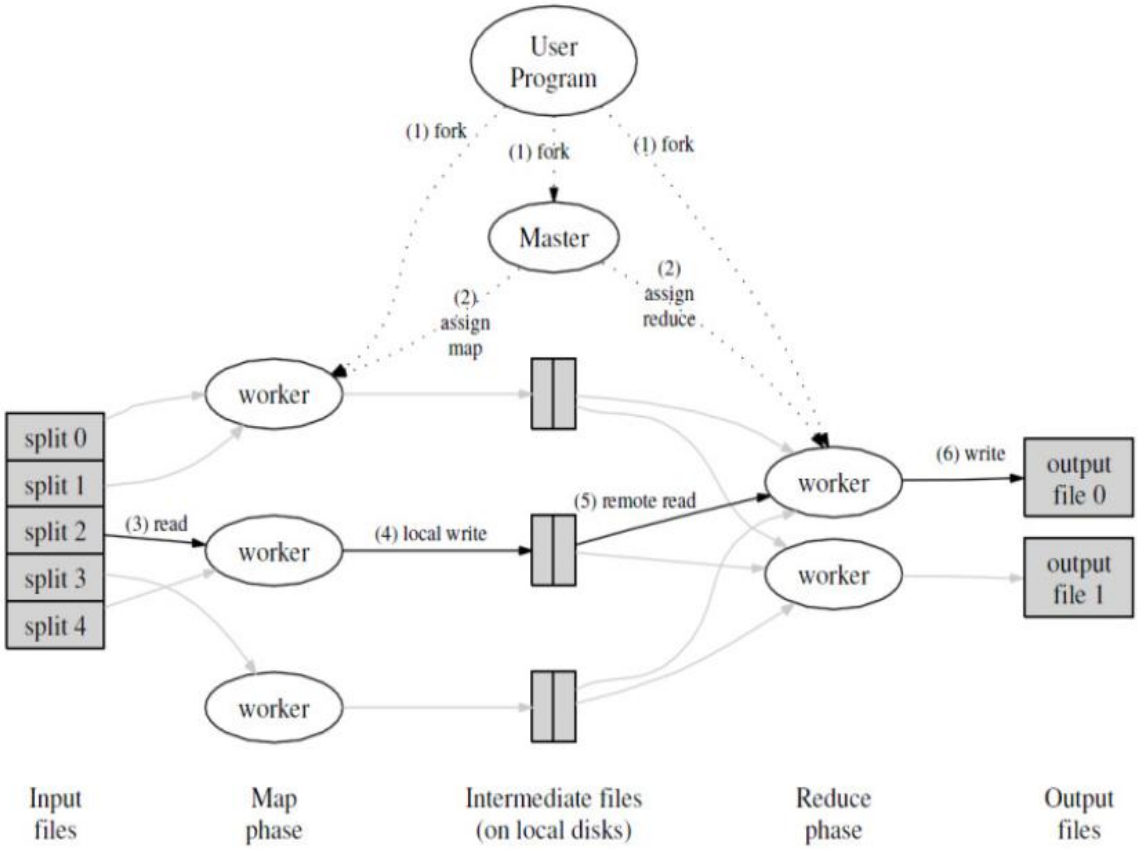
流处理



批处理

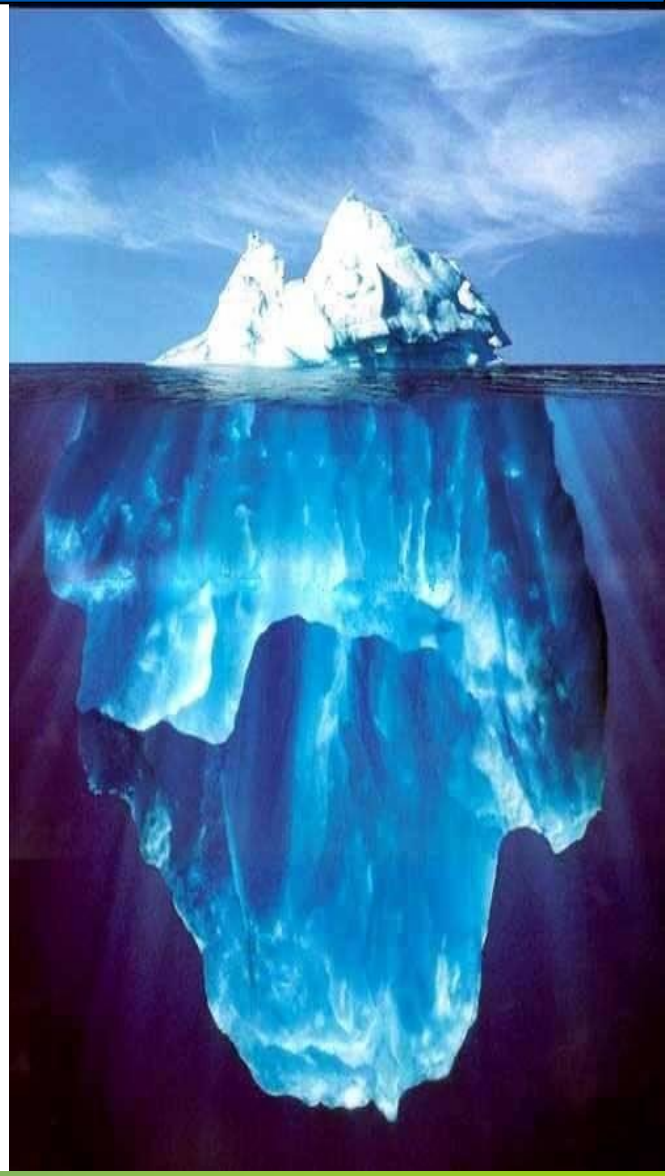
Google 2004 MapReduce

MapReduce



提纲

- 大数据处理的基本流程
- 大数据处理模型
- 大数据关键技术
- 大数据处理工具
- 大数据时代面临的新挑战



大数据关键技术

Google 2006

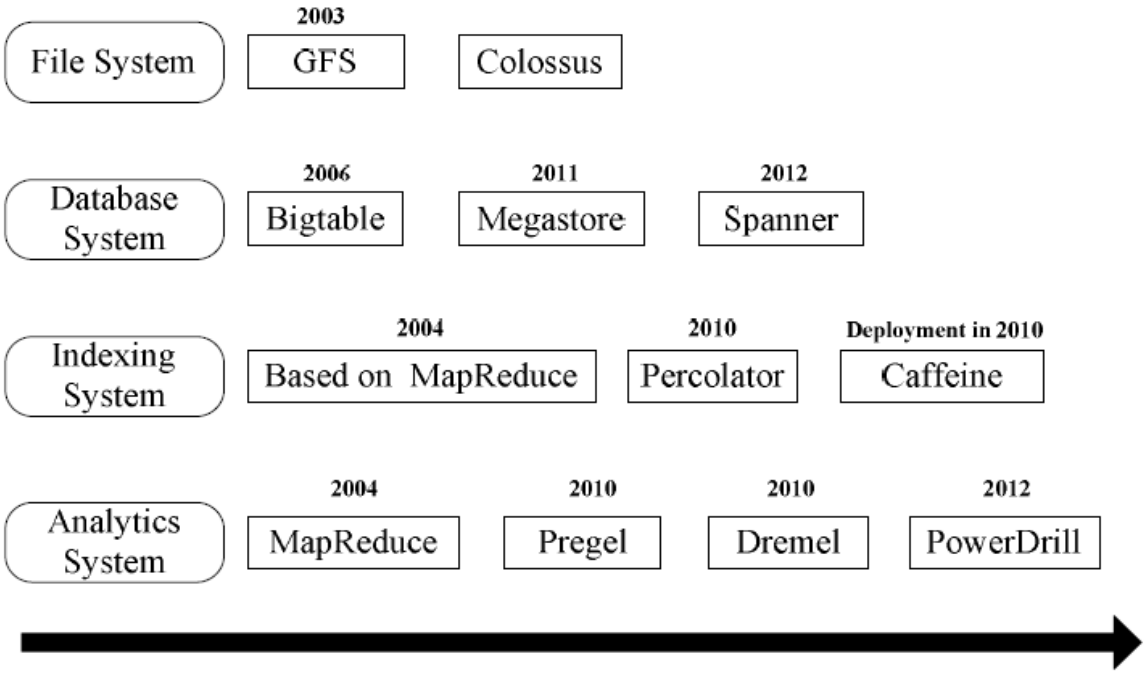
Google

GFS

MapReduce Bigtable

Hadoop

Google



文件系统

Google Facebook

- GFS
- Colosuss
- HDFS
- CloudStore
- Haystack
- TFS
- FastDFS

数据库系统

1.
2.
3.
4.

Google

1. Google Bigtable
2. Amazon Dynamo
3. Yahoo PNUTS

NoSQL技术

Bigtable Dynamo PNUTS

NoSQL(Not Only SQL)

NoSQL NoSQL

1. (schema-free)
2. (easy replication support)
3. (simple API)
4. (BASE ACID)
5. (Huge amount of data)

索引和查询技术

Facebook

Algorithms	Index Size for Facebook	Index Time for Facebook	Query Time on Facebook(s)
Ullmann[Ullmann 76]	-	-	>1000
VF2[CordellaFSV04]	-	-	>1000
RDF-3X[NeumannW10]	1T	>20 days	>48
BitMat[AtreCZH10]	2.4T	>20days	>269
Subdue[HolderCD94]	-	>67 years	-
SpiderMine[ZhuQLYHY11]	-	>3 years	-
R-Join[ChengYDYW08]	>175T	$>10^{15}$ years	>200
Distance-Join[ZouCO09]	>175T	$>10^{15}$ years	>4000
GraphQL[HeS08]	>13T(r=2)	>600 years	>2000
Zhao[ZhaoH10]	>12T(r=2)	>600 years	>600
GADDI[ZhangLY09]	$>2*10^5$T(L=4)	$>4*10^5$ years	>400

索引和查询技术

NoSQL

NoSQL NoSQL

1. **MapReduce** MapReduce
   NoSQL MapTask

2.
   NoSQL

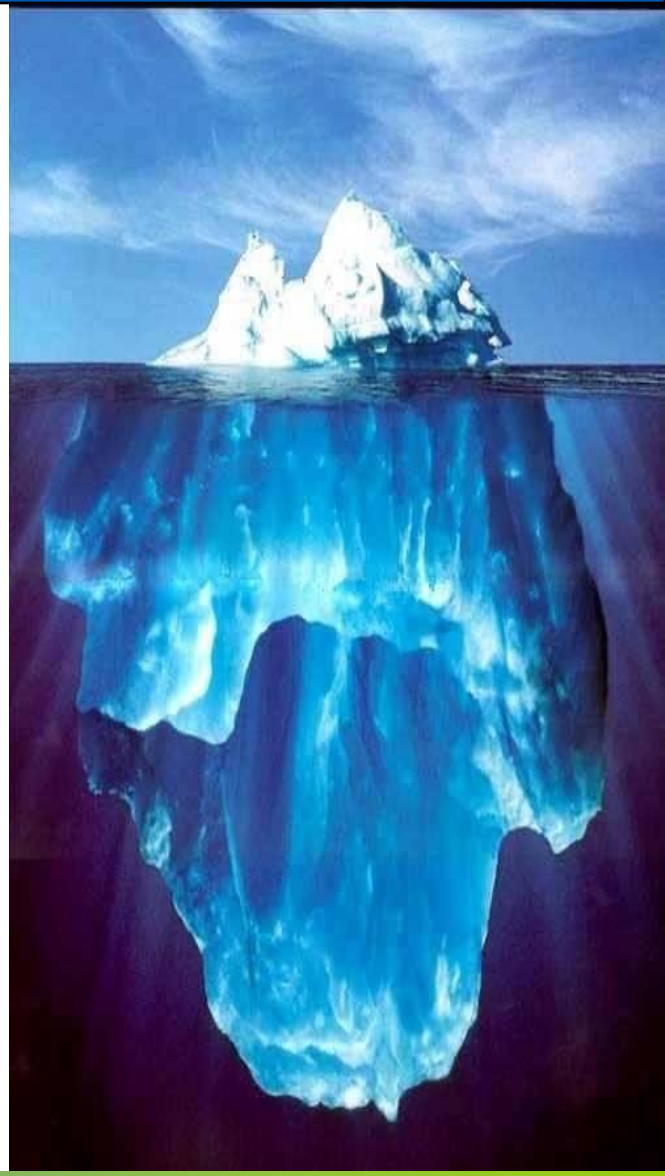
数据分析技术

- Mapreduce
- Pregrel
- Dremel Web
- PowerDrill

  - Storm
  - Percolator\Nectar\DryadInc
  -

提纲

- 大数据处理的基本流程
- 大数据处理模型
- 大数据关键技术
- 大数据处理工具
- 大数据时代面临的新挑战



大数据处理工具

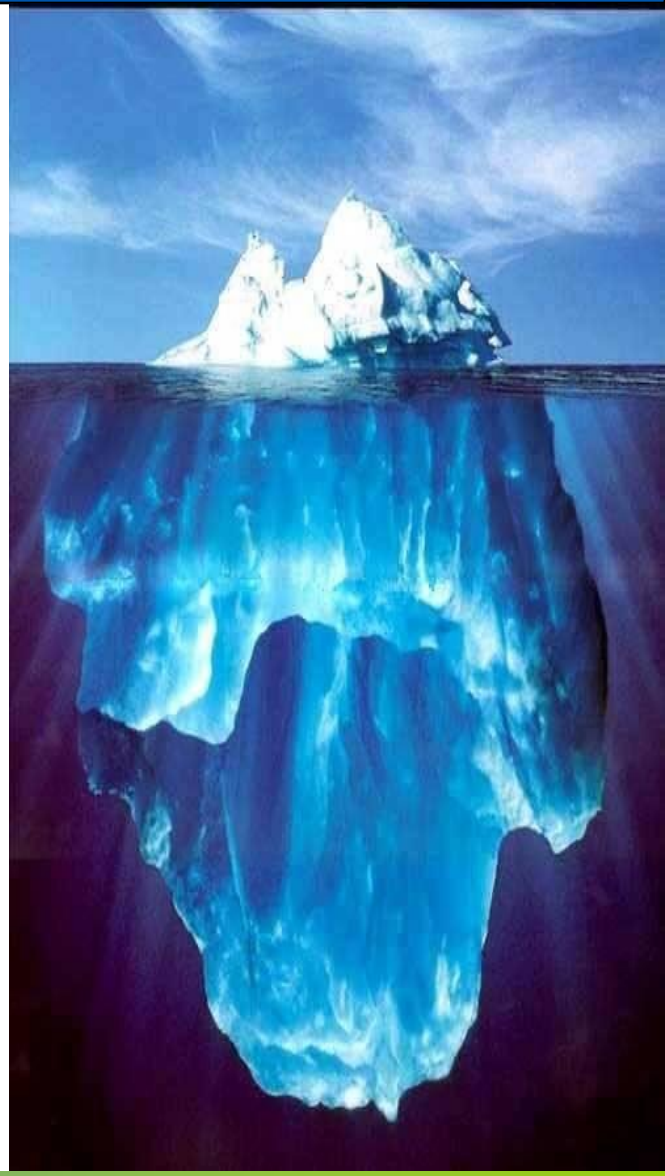
Hadoop

Hadoop

Category		Examples
Platform	Local	Hadoop、MapR、Cloudera、Hortonworks、InfoSphere BigInsights、ASTERIX
	Cloud	AWS、Google Compute Engine、Azure
Database	SQL	Greenplum、Aster Data、Vertica
	NoSQL	HBase、Cassandra、MongoDB、Redis
	NewSQL	Spanner、Megastore、F1
Data Warehouse		Hive、HadoopDB、Hadapt
Data Processing	Batch	MapReduce、Dryad
	Stream	Storm、S4、Kafka
Query Language		HiveQL、Pig Latin、DryadLINQ、MRQL、SCOPE
Statistic and Machine Learning		Mahout、Weka、R
Log Processing		Splunk、Loggly

提纲

- 大数据处理的基本流程
- 大数据处理模型
- 大数据关键技术
- 大数据处理工具
- 大数据时代面临的新挑战



大数据集成

-
  -
  -
  -
-
  -

大数据分析

- (Timeliness)
-
-

大数据隐私问题

-
-
-

大数据能耗问题

大数据处理与硬件的协同

-
  - Mapreduce
-

大数据管理易用性问题

-
-
- (Visibility)
- (Mapping)
- (Feedback)

性能的测试基准

-
-
-
-
-

本章小结

Google

Google

Hadoop

参考文献

- [1] , . . , 2013 8 .

主讲教师和助教

主讲教师：林子雨

E-mail: ziyulin@xmu.edu.cn
http://www.cs.xmu.edu.cn/linziyu
http://dblab.xmu.edu.cn

助教：赖明星

2011

E-mail: mingxinglai@gmail.com
http://mingxinglai.com

2013 http://dblab.xmu.edu.cn/node/423

PPT

Department of Computer Science, Xiamen University, Sep, 2013