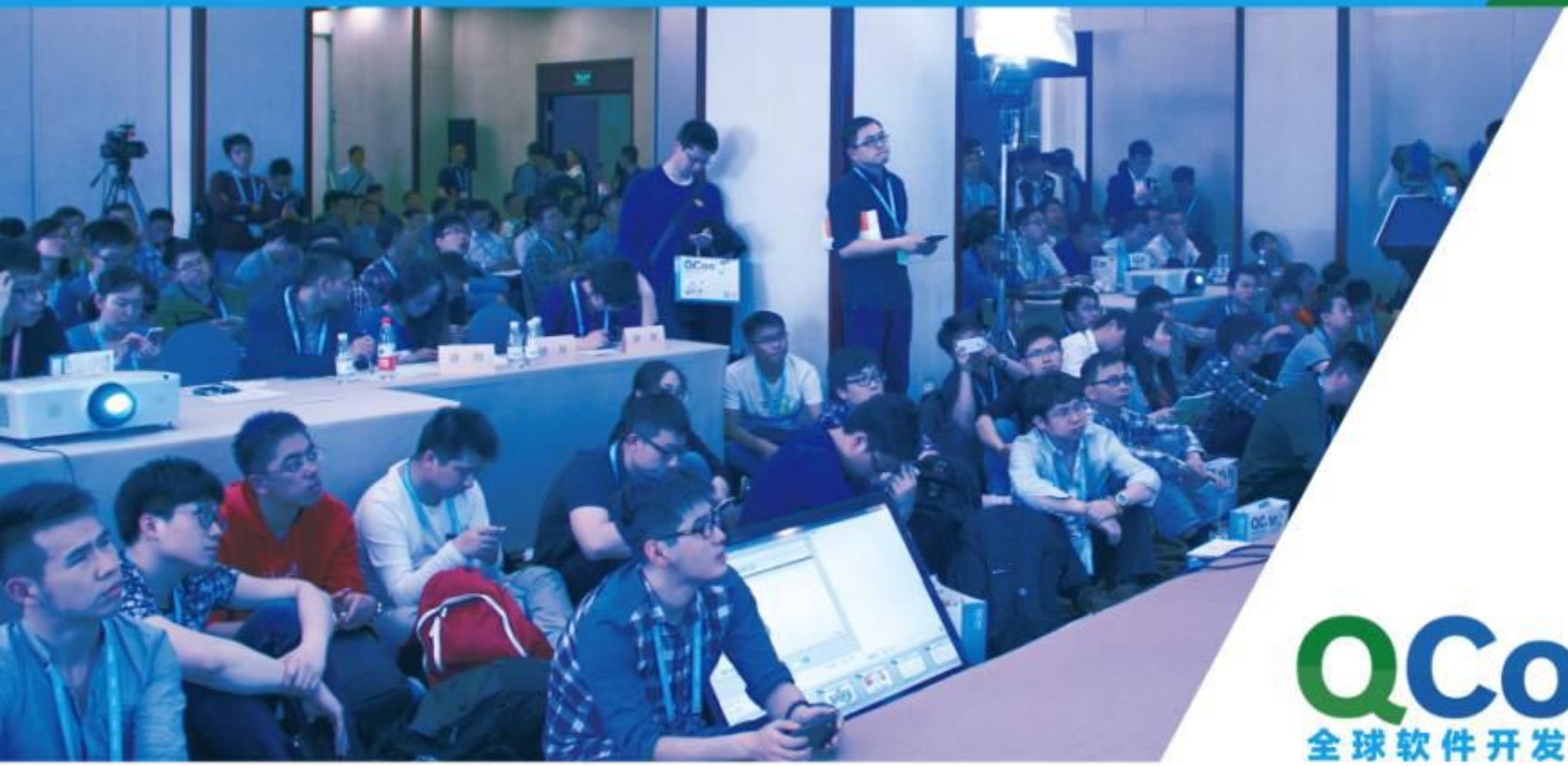
# QCon全球软件开发大会

International Software Development Conference



QCon
全球软件开发大会

# 开源大数据在Facebook与Dropbox的实践

邵铮
前Dropbox/Facebook研发经理
Apache Hadoop PMC成员

# 关于我自己

- 时间：职业经历 [开源软件]
- 2005-2008：Senior Software Engineer, **Yahoo Web Search [Hadoop]**
- 2008-2014：Senior Engineering Manager, **Facebook**
  - 2008-2010：Staff Software Engineer, **Data Infrastructure [Hive]**
  - 2010-2012：Engineering Manager, **Data Freeway [Scribe]**
  - 2012-2014：Senior Eng Manager, **Database Engineering [MySQL, RocksDB]**
- 2014-2015：Engineering Manager, **Dropbox [MySQL, Hive, Presto, Scribe, RocksDB]**
- 2015-Now: Senior Engineering Manager, **Data Infra, Uber [a lot more]**

# 关于我自己 (2)

- 2008-2012: Apache Hive PMC Member
  - Hive创始团队成员
- 2009-Now: Apache Hadoop PMC Member
  - HDFS, Map-Reduce的Committer
- 2013-2014: 支持我的团队开发和开源RocksDB
- Now: 关注开源项目的社区、生态、与创业

# 关于我自己 (3)

- 开源软件的受益者
- 开源软件的贡献者
- 开源软件的信徒

# 开源软件的作用

- 对个人：
  - 增强Impact, Influence
  - 促进职业发展，有机会直接成为创始人
- 对公司：
  - 最大化投资回报率(ROI)
  - 招聘与长期维护
- 对社会：
  - 提高生产效率，减少重复的轮子
  - 增进优胜劣汰

# 下面的章节

- 数据仓库
- 流数据处理
- NoSQL
- Dropbox的开源大数据策略
- 展望未来

# 数据仓库

# Hadoop的诞生

- 2006年，Yahoo 为什么投资开发Hadoop?
  - 与Google的搜索引擎大战
  - 薄弱的基础架构急需更新
- Yahoo为什么用开源的方式来开发Hadoop?
  - Doug Cutting
  - 行业第二的最优策略是联合所有人来挑战行业第一

# Hadoop的先天缺陷

- 接口：底层接口优先
  - Map-Reduce功能强大但是不容易使用
  - Cascading, Pig是为engineer和scientist设计的
- 性能：为Big Job而优化，忽略了Small Job的overhead
  - Schedule时间, JVM启动时间, 1秒/数秒一次的heartbeat
  - 可扩展性大大优于性能

# Facebook的数据处理需求

- 2007-2008年的状况
  - 原始数据量急速上升
  - Oracle RAC
- 超大规模数据仓库的需求
  - 长期可以支持Facebook 2008年数据量的1000x
  - 使用对象是engineer, scientist, 和analyst (both technical and non-technical)
- 商用数据仓库无法满足如此庞大的scale
  - TeraData, Netezza (IBM), Aster Data (TeraData), GreenPlum (EMC)

# Hive的诞生

- 为什么Facebook要开发Hive?
  - CTO Adam D'Angelo非常重视data-based decision
  - 决不能让系统的scalability限制了业务的发展
- 为什么基于Hadoop开发Hive?
  - 站在巨人的肩膀上，直接解决了scalability的问题
  - Hadoop的开源社区非常强大
- Hive最大的创新在哪里?
  - SQL on Hadoop

# Hive与Pig之争

- Pig Team：
  - 我们早就有计划要做SQL。你们为什么不在Pig基础上做SQL?
- Hive Team：
  - 你们的代码性能太差，我们等不及。
- “Does it really matter to reinvent the wheel?”
  - No!  As long as your project wins.
- “The best way to work in open-source is to compete to death.” – a senior leader in the Hadoop ecosystem

# Hive发展过程中的重要决定

- 发行
  - 作为Hadoop的子项目发行
  - 升级为Apache顶级项目
- 推广
  - 各大会议的演讲 (Hadoop Summit, Hadoop World, Hadoop in China)
  - Papers – 获得学校和研究院的支持 (Berkeley, Yale, OSU, 中科院计算所等等)
  - Meetup，各大公司的合作 (AWS, Netflix, Taobao, etc)
- 即装即用
  - 内置Derby数据库
  - 邮件列表/JIRA的支持

# Hive简要架构

HDFS
Tachyon

Map-Reduce
Tez
Spark

Hive CLI / Beeline
Browse / Query / DDL

Web UI
Management

Hive QL
Parser/Planner
Optimizer/Execution

Hive Metastore Server
Thrift

FileFormat
/SerDe
RC/ORC/Parquet

Hive Metastore
Derby / MySQL

# Hadoop数据仓库图解

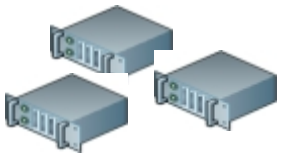
Business intelligence
[Tableau, Microstrategy]

Web UI / CLI / Queries
[Hive/Spark/Zeppelin]

Analytics for users
[HighCharts, D3]

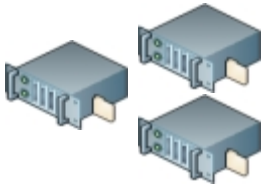
Data Mart
[Vertica, MemSQL]

**Hadoop Data Lake**



Database

Workflow Management
[Oozie, AirFlow, Pinball]



Log / Streaming Data
[Kafka, Flume]

Web/Mobile Servers

# 新一代开源大数据架构

| [开源软件] | 商业分析平台 | 产品分析平台 | 工程分析平台 |
|---|---|---|---|
| | 公司仪表盘 | A/B Test平台 | 推荐系统 |
| | 商业智能 BI<br>*Tableau, Microstrategy* | 多维度分析平台<br>[Kylin, Presto] | 机器学习<br>[MLLib, H2O, Weka] |
| | 数据超市Data Mart<br>*Vertica,* [MemSQL] | SQL分析平台<br>[AirPal, Zeppelin, HUE] | 社交、位置分析<br>[Giraph, Hadoop-GIS] |
| 工作流 | [Airflow, Chronos, Pinball, Oozie] | | |
| **数据建模** | 商业、产品、工程的数据定义，如Revenue, MAU, CTR | | |
| **元数据** | 库/表/列/注释/静动态关系/负责人等，存储于[Hive Metastore] | | |
| 数据湖 | [Hive, Spark, Impala, Pig, Cascading, Tez] + [HDFS + YARN/Mesos] | | |
| 数据导入 | [Gubblin/Camus, Sqoop, etc] | | |
| 数据源 | Log [Kafka/Flume, ActiveMQ] | Database [MySQL, NoSQL etc] | |

# Hive使用中出现的三大问题

- 数据延迟太大
  - 流数据处理
- 查询延迟太大
  - 交互式查询
- Map-Reduce框架影响了性能
  - 新的框架：Spark RDD，Apache Tez

# Presto – Facebook对Hive的自我革命

- Hive无法满足交互式查询的速度要求
  - 继续improve Hive，还是另起一个项目？
- Peregrine (Presto前身)
  - 一个人的项目验证可行性
- Presto
  - 基于Google Dremel的设计
  - Facebook的大力支持
  - 目前广泛应用于Dropbox, Airbnb等公司

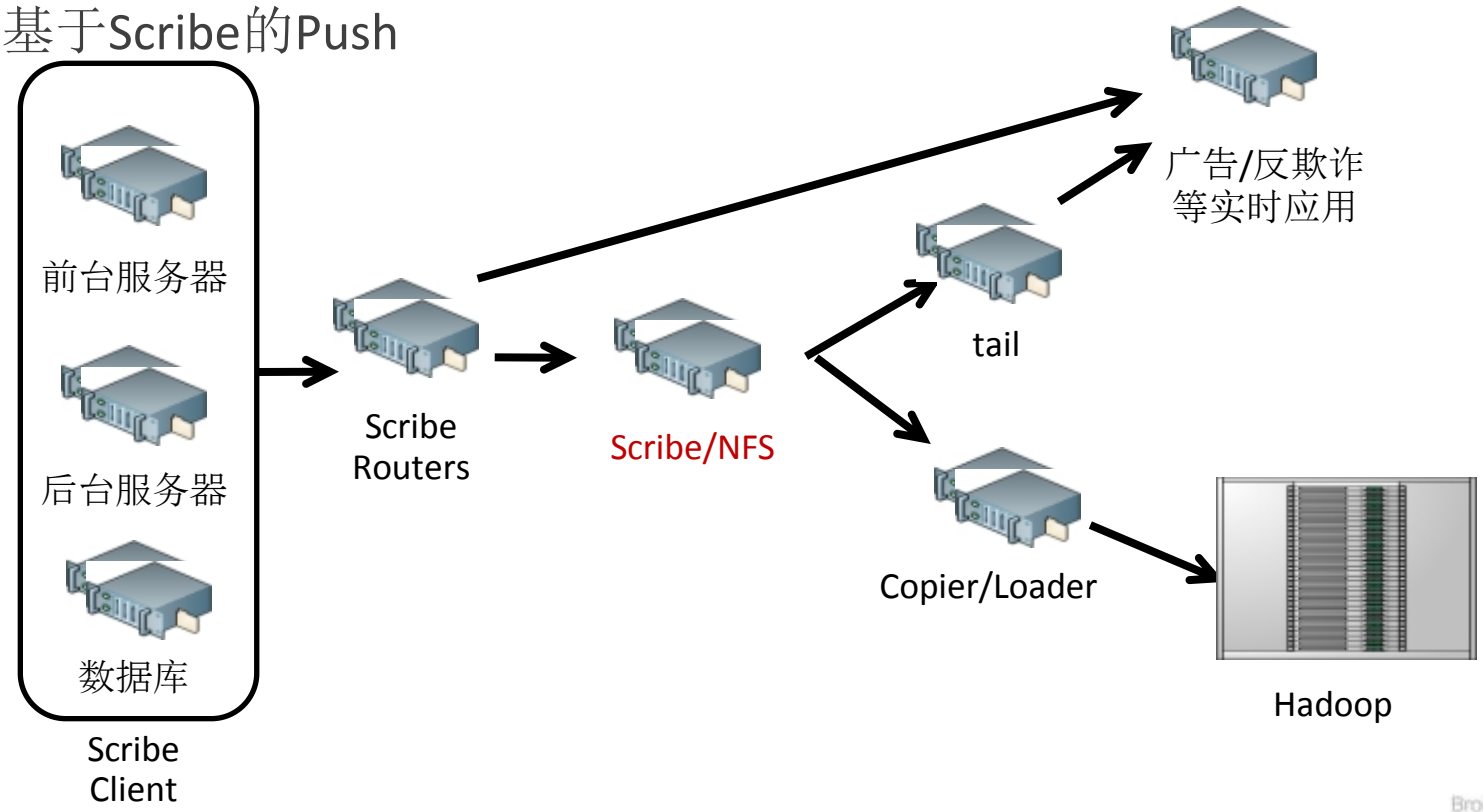
# 流数据处理

# Facebook的流数据处理需求

- 关键应用日益庞大
  - Site Integrity – Anti-Spam与Safety
  - Analytics – Ads与Developer Platform
- 原有架构的Scale无法跟上数据量乘积式的上升
  - 大规模的用户增长
  - 产品的复杂度与公司员工的数量
- Data Freeway的诞生
  - 2012年达到数十GB/秒的低延迟(10秒级别)的流数据传输和处理系统
  - （这个流量大约是Kafka 2015年在LinkedIn的规模的十倍）

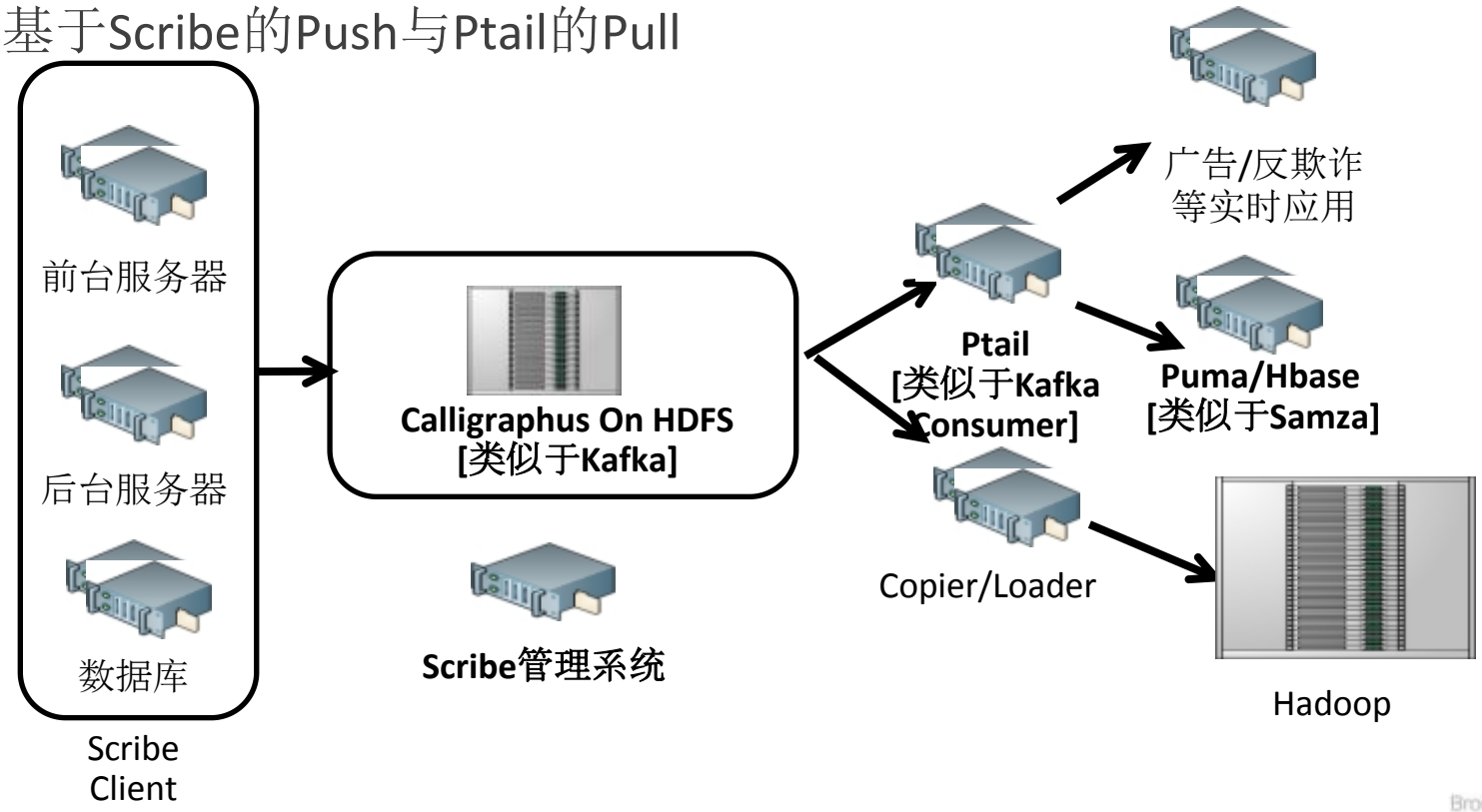
# Data Freeway流处理架构图 (1)

- 基于Scribe的Push



# Data Freeway流处理架构图 (2)

- 基于Scribe的Push与Ptail的Pull



# Data Freeway的发展

- 在Facebook：成功
  - 数年来稳定地承载着Facebook的巨大数据流量
  - 支撑着越来越多的应用场景

- 在业界：失败
  - 由于没有Open-source，在业界没有任何的影响力
  - 几年之后，Storm, Kafka, Samza等系统脱颖而出，占据了主流
  - 原Kafka团队脱离LinkedIn成立了startup: Confluent.io

# 数据流处理下一步的挑战

- 与整个数据仓库架构的整合
  - 数据流能不能成为数据湖的一部分？
  - 数据流和数据湖能不能用同一套架构来处理数据？
- 迟到的数据的处理
  - 如何知道当前的数据已经全了？
  - 迟到的数据出现时，如何让用户选择重新处理还是忽略？

# NoSQL

# HBase与Cassandra

- Cassandra – 基于Amazon Dynamo
  - Facebook开发，用于Facebook Inbox的后端存储。
- HBase – 基于Google BigTable
  - 社区推动，2010年Facebook决定用于Facebook Message (Inbox下一代产品) 的后端存储。同时，Facebook放弃了Cassandra的继续开发。
- 为什么Facebook选择了HBase?
  - Google设计的架构，应该不会错
  - HBase的社区发展很快

# HBase可以取代MySQL吗？

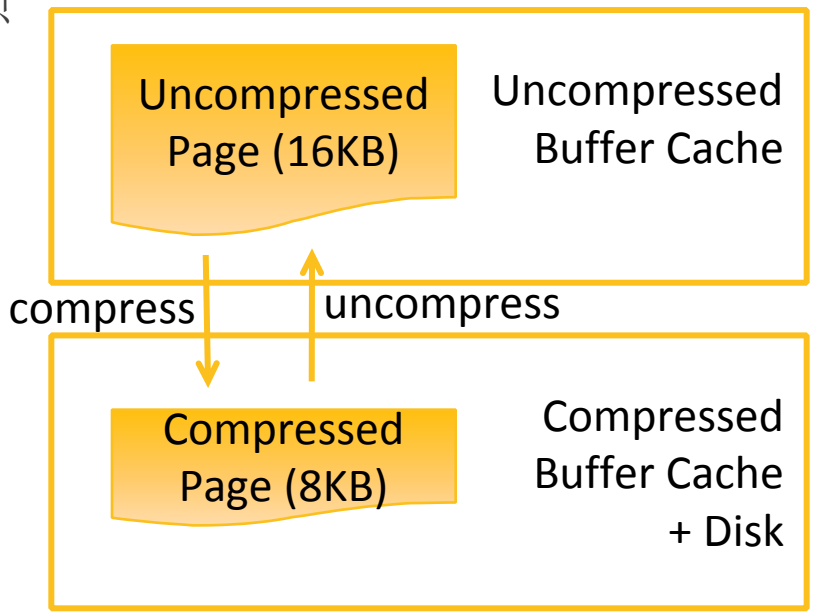
- MySQL的弱点
  - 没有原生的Cluster Management: sharding, replication需要手动配置
  - InnoDB的B-Tree结构： Optimized for read, not write
  - MySQL的瓶颈在write IO上。
- HBase的优势
  - 原生的Cluster Management支持
  - LSM的结构Optimized for write
  - 读的效率低，但是可以由caching layer来解决
- 但是…

# Hbase没能取代MySQL

- Facebook投入了4个工程师2年时间，试图用HBase取代MySQL，结果项目取消。
- 原因：
  - HBase的成熟程度远不及MySQL。需要大量的tuning。
  - SSD的普及使得IO不再是瓶颈。瓶颈转移到了CPU。
  - MySQL在同期有了大幅度的提升。

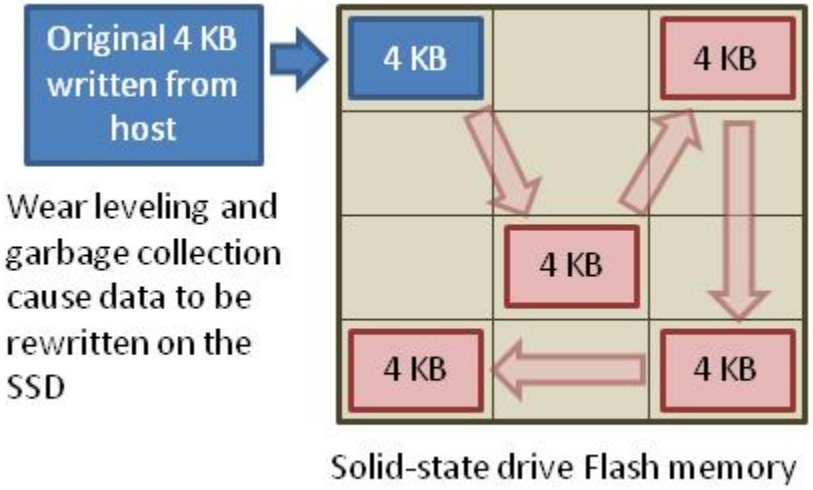
# MySQL的最大提升： Compression

- SSD给MySQL带来的机会
  - 原本IO是瓶颈，现在存储空间变成了瓶颈
- Compression
  - 用8KB的空间来存储16KB的Page
  - 每次读写的时候compress和decompress
  - CPU出现瓶颈
  - Lazy compression的优化



# 什么才是最适合SSD的数据结构？

- SSD的弱点：
  - Sometimes slow
    - No in-place write
    - Wear Leveling
    - Garbage Collection
  - Expensive
- WAF (Write Amplification Factor)
  = physical writes / logical writes
- 目标：降低WAF，加大压缩比

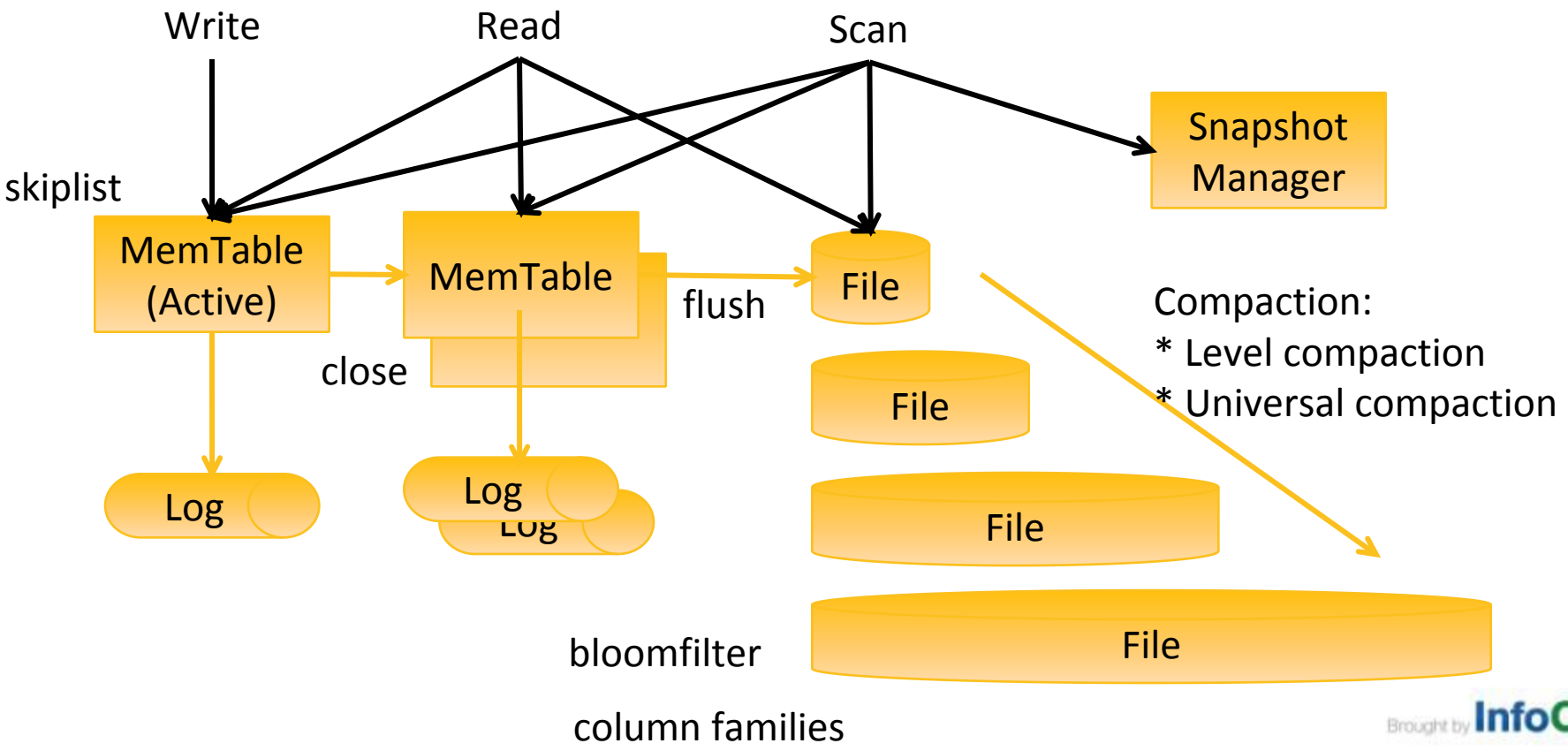



source: wikipedia

# RocksDB的诞生

- HBase项目的原班人马，寻找最适合SSD的存储引擎
- RocksDB
  - 基于Google Open-Source的LevelDB (LSM数据结构)
  - 针对SSD/Memory做了大量的性能优化
  - 作为Embedded database，直接应用在Facebook Newsfeed后端
  - Open-Source: http://rocksdb.org

# RocksDB的架构


Write
Read
Scan
Snapshot Manager
skiplist
MemTable (Active)
MemTable
flush
File
close
File
Log
Log
File
Compaction:
* Level compaction
* Universal compaction
bloomfilter
File
column families

# RocksDB的现状

- CockroachDB – RocksDB-powered Geo-Replicated Transactional Database
- MyRocks – RocksDB-powered MySQL (Development Mode)
  - 降低CPU costs
  - 实现pessimistic concurrency control
  - 降低P99 read latency
  - Port ZSTD compression
- MongoRocks - RocksDB-powered MongoDB (Production ready)
  - 提高社区接受度
  - 支持新的MongoDB 3.2版本

# Dropbox的开源大数据策略

# 技术策略

- 早期紧跟Facebook
  - 2011年上线Scribe, Hive
  - 2013年底上线Presto
- 现在紧跟社区
  - 2014年上线Kafka，重写Scribe
  - 2015年升级到YARN （Facebook的Corona一直没有开源）
  - 2015年测试SparkSQL/SparkStreaming
- 社区选择太多？
  - 兼顾 – 开发SQL Translator，支持Presto, Hive, SparkSQL

# 人才策略

- 联合各大公司合作与交流
  - 社区专家 – Percona, Cloudera
  - 同类公司 – 湾区各大Unicorn公司
- 大力招聘开源社区的专家
  - 使用开源软件大部分的时间都在填坑
  - 减少跳坑的可能
- 大力培养开源软件的熟手
  - 开源社区专家的稀缺
  - 自己培养丰衣足食

# 可以改进的机会

- 没有尽可能的利用社区已有的项目
  - OpenTSDB
- 改进了社区的项目但是没有开源出来
  - Goscribe – A much better scribe
- 开源出来了但是没能进入主流
  - Presto-Kafka Connector

# 总结与展望

# 开源的趋势

- 非开源的技术会逐渐被开源取代
  - Hadoop vs commercial competitors
  - Kafka vs Scribe
  - Storm/Heron/Samza vs Puma
- 新的开源技术会一直层出不穷
  - Spark, Tachyon, Flink, etc
  - 但是中国自主创新的国际流行的开源项目非常少

# 开源 * 学校

- 开源是最好的学习机会
  - 学校的课程设计要大力融合开源软件的实践
- 开源社区
  - 学生科协组织兴趣小组，一起融入开源社区
- 开源比赛
  - 从Topcoder类型的娱乐性比赛转向开源软件的公益性比赛

# 开源 * 服务平台

- 开源社交平台
  - 目前的开源平台都是工具平台，而不是社交平台
- 开源众包和招聘平台
  - 连接顶级公司与开源领域的人才

# THANKS

http://www.linkedin.com/in/zshao
微信: zshao9