



# 大众点评网

## 监控平台剖析

吴其敏 @QCon Hangzhou 2012  
[qimin.wu@dianping.com](mailto:qimin.wu@dianping.com)

# 大众点评网

- 2003年4月成立于上海，是中国领先的本地生活消费平台
- 覆盖全国2300座城市，180万家商户，2200万条消费者点评
- 月活跃用户数5400+万，移动客户端独立用户数4500+万
- 500+台应用服务器，300+名技术人员

(数据截止2012年第三季度)

# Agenda

- 介绍
- 报表
- 设计
- 未来

# 监控系统

- MRTG, Cacti, Nagios, Ganglia, Zabbix, ...
- Dapper(Google), Scribe(Facebook), Zipkin(Twitter), CAL(eBay), ...
- CAT(Dianping)

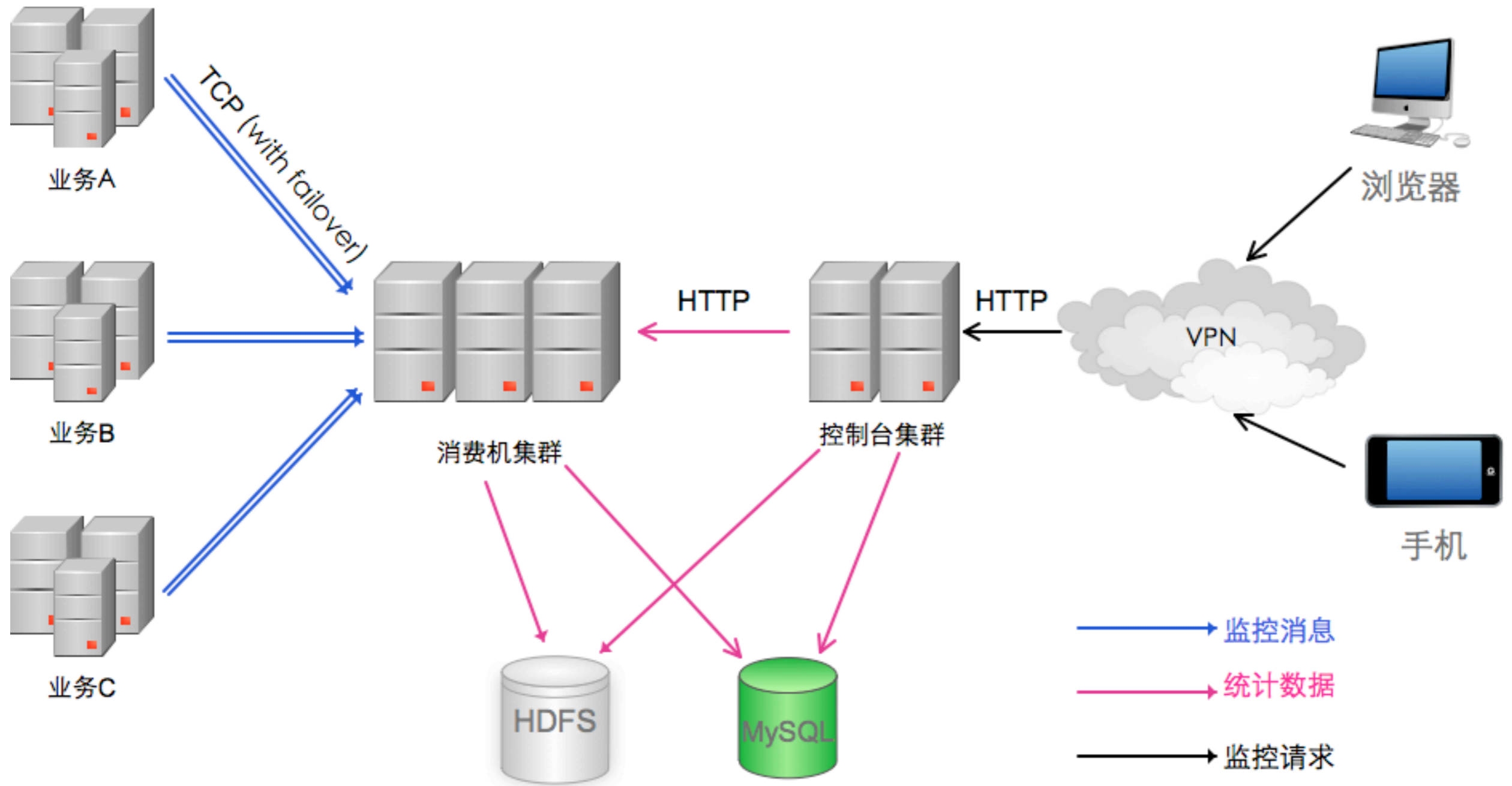




# 目标

- 故障诊断
  - 跨越边界访问(Across Boundary Activity)
    - 跨网络: URL, HTTP, RPC, SQL, Message, Cache, ...
    - 跨角色: 浏览器/网络/应用服务器, Nginx/Tomcat, Controller/Model/View, ...
    - 跨语言: Java call Scala, PHP call Java, ...
    - 跨所有权: 公共组件, 规则引擎, ...
  - 状态记录 (Status)
    - 系统状态: CPU, Memory, Thread, Disk, Network, GC, Load, ...
    - 业务状态: 参数, 结果, ...
- 业务统计
  - 计数/计时: Page, Service, Cache, Database
  - 分布: 浏览器, IP, 登陆状态, 缓存命中, ...
  - A/B 测试, 性能测试, ...
- 系统优化: 响应时间, 依赖耦合, 嵌套调用
- 容量规划: 数据库, 应用服务器

# 部署



# 特性

- 实时处理，全量数据，侧重应用监控
- 轻量级，低开销
- 无中心化设计，支持水平扩展
- 通用API，支持各种业务
- 丰富的分析、统计报表



# 目前

- 集成所有中间件产品
- 60+ 业务应用
- 400+ 应用服务器
- ~1TB 消息（每天）
- ~200GB 存储（每天，压缩后）

# Agenda

- 介绍
- 报表
- 设计
- 未来

# 消息



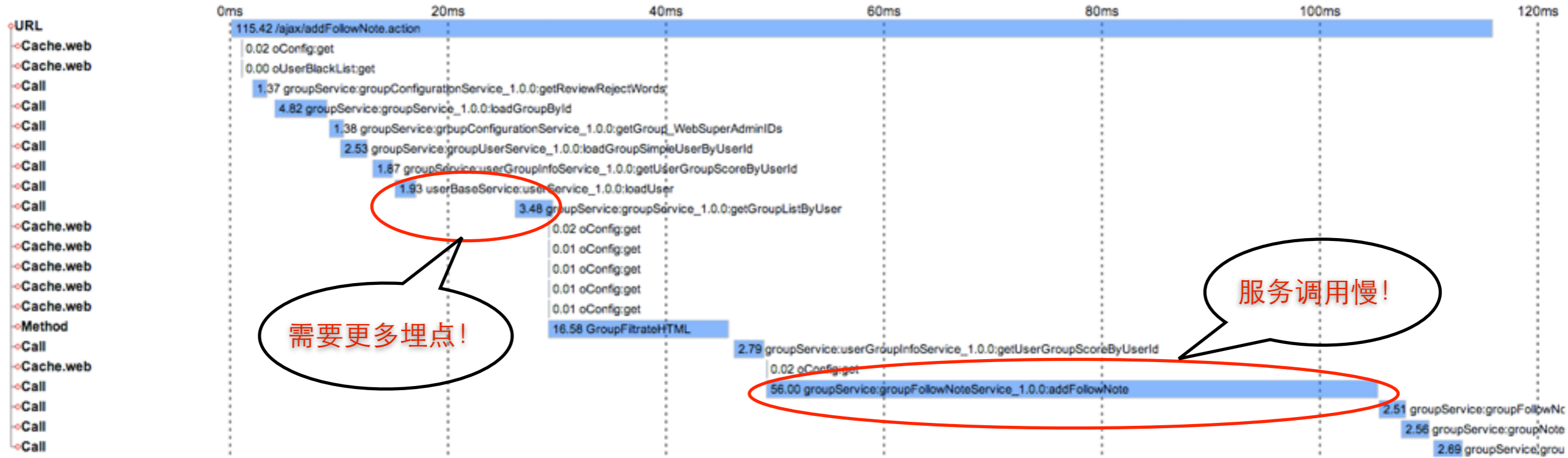
- 消息头
  - 版本号，消息ID，所属业务，IP，所在线程，根消息ID
- 消息体

Type & timestamp	1st Category	2nd Category	Status	Duration & Attributes
t14:38:56.595	URL	t		
E14:38:56.595	URL.Server	cat.dianpingoa.com		RemoteIP=[redacted]&Referer=http://cat.dianping
E14:38:56.595	URL.Method	HTTP/GET		/cat/r/t?domain=&date=2012101314&reportType=
A14:38:56.595	MVC	InboundPhase		0.06ms
A14:38:56.595	MVC	TransitionPhase		0.00ms
t14:38:56.595	MVC	OutboundPhase		
t14:38:56.595	ModelService	CompositeTransactionService		
A14:38:56.596	ModelService	RemoteTransactionService		1.06ms http://[redacted]:8080/cat/r/model/transact
A14:38:56.596	ModelService	RemoteTransactionService		0.86ms http://[redacted]:8080/cat/r/model/transac
A14:38:56.596	ModelService	RemoteTransactionService		1.89ms http://[redacted]:8080/cat/r/model/transac
A14:38:56.596	ModelService	RemoteTransactionService		1.79ms http://[redacted]:8080/cat/r/model/transac
A14:38:56.596	ModelService	RemoteTransactionService		27ms http://[redacted]:8080/cat/r/model/transacti
T14:38:56.622	ModelService	CompositeTransactionService		27ms request=ModelRequest[domain=Cat, period
T14:38:56.628	MVC	OutboundPhase		33ms
T14:38:56.628	URL	t		33ms module=r&in=t&out=t

t: Transaction Start  
E: Event  
T: Transaction End  
A: Atomic Transaction

Transaction: 可嵌套  
Event: 不可嵌套  
Heartbeat: 不可嵌套

# 消息



# 消息树

t15:00:44.023	URL	/ajax/addVote.action	
E15:00:44.023	URL	ClientInfo	RemotelP=180.175.162.12
E15:00:44.023	URL	Payload	HTTP/POST /ajax/addVote
A15:00:44.023	Cache.web	oConfig:get	0.02ms finalKey=oConfig
A15:00:44.023	Cache.web	oUserBlackList:get	0.00ms finalKey=oUserBl
t15:00:44.026	Call	groupService:groupSurveyService_1.0.0:addVote	
	<a href="#">[:: hide ::]</a>		
t15:00:43.967	Service	groupService:groupSurveyService_1.0.0:addVote	
E15:00:43.967	PigeonRequest	Payload	
t15:00:43.967	SQL	GroupSurvey.loadSurvey	
E15:00:43.967	SQL.Method	Select	
E15:00:43.968	SQL.Database	jdbc:mysql:// [REDACTED] ?characterEncoding=U	
T15:00:43.967	SQL	GroupSurvey.loadSurvey	
t15:00:43.968	Call	userBaseService:userService_1.0.0:loadUser	
	<a href="#">[:: hide ::]</a>		
t15:00:44.089	Service	userBaseService:us	
E15:00:44.089	PigeonRequest	Payload	
A15:00:44.089	Cache.memcached	eUserAtUC:get	
T15:00:44.089	Service	userBaseService:us	
	<a href="#">[:: show ::]</a>		
T15:00:43.970	Call	userBaseService:userService_1.0.0:loadUser	
A15:00:43.970	Cache.memcached	oUserGroupScore:get	
t15:00:43.975	SQL	GroupSurvey.addVote	
E15:00:43.975	SQL.Method	Execute	
E15:00:44.244	SQL.Database	jdbc:mysql:// [REDACTED] ?characterEncoding=U	
T15:00:44.243	SQL	GroupSurvey.addVote	
T15:00:44.244	Service	groupService:groupSurveyService_1.0.0:addVote	
	<a href="#">[:: show ::]</a>		
T15:00:44.305	Call	groupService:groupSurveyService_1.0.0:addVote	279ms CallType=sync
T15:00:44.307	URL	/ajax/addVote.action	284ms

# Transaction 报表

Machines: [ All ] [ ] [ ] [ ] [ ] [ ] [ ]

Type	Total Count	Failure Count	Failure%	Sample Link	Min(ms)	Max(ms)	Avg(ms)	95Line(ms)	Std(ms)	TPS
URL	241,769	0	0.00%	<a href="#">Log View</a>	0.3	3543.4	162.9	513.0	195.3	77.4
System	238	0	0.00%	<a href="#">Log View</a>	14	73.5	23.6	35.0	5.9	0.1
Method	3,026	0	0.00%	<a href="#">Log View</a>	9.6	296.8	21.2	31.8	20.2	1.0
SQL	54	0	0.00%	<a href="#">Log View</a>	4.4	70.2	19.3	58.4	17.1	0.0
Call	4,249,777	0	0.00%	<a href="#">Log View</a>	0.1	3299.7	6.9	16.4	39.0	1359.9
Cache.memcached	95,171	0	0.00%	<a href="#">Log View</a>	0.1	40.2	1.1	3.0	1.3	30.5
Result	4,211,145	0	0.00%	<a href="#">Log View</a>	0	112.3	0.2	0.0	0.3	1347.6
Cache.web	40,998,534	0	0.00%	<a href="#">Log View</a>	0	46	0.0	0.0	0.1	13119.4



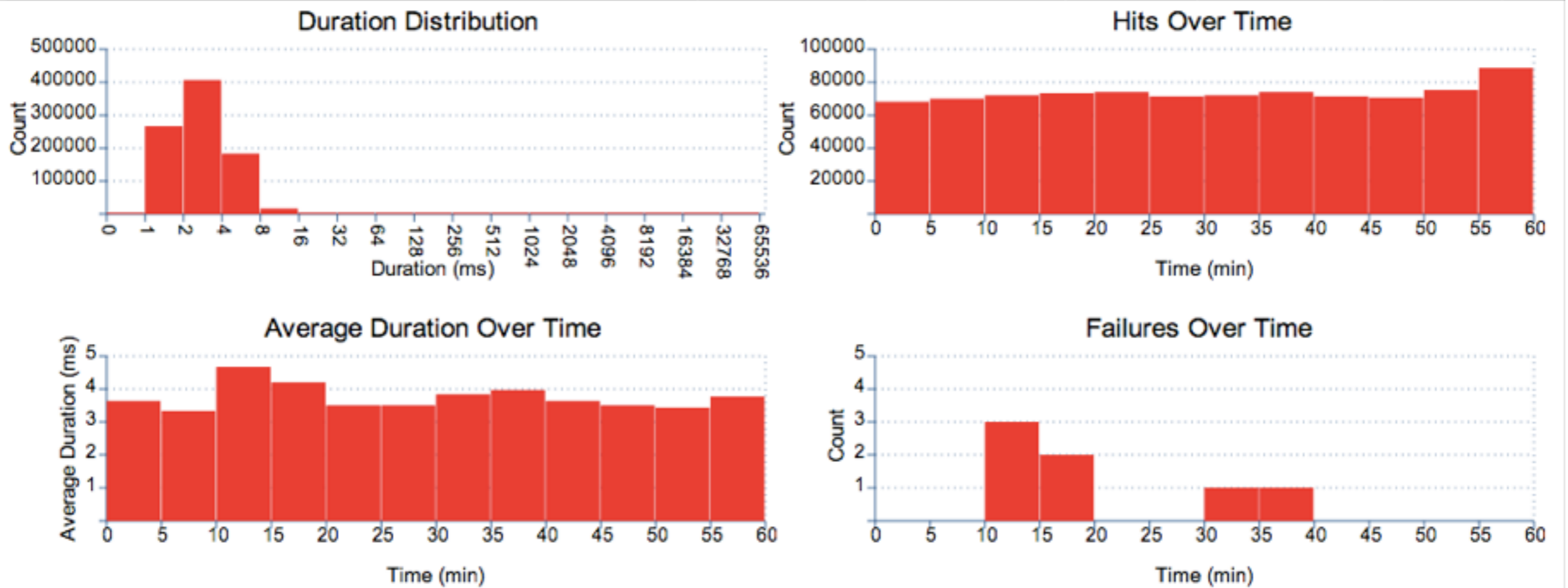
# Transaction 报表

Machines: [ All ] [ ] [ ] [ ] [ ] [ ]

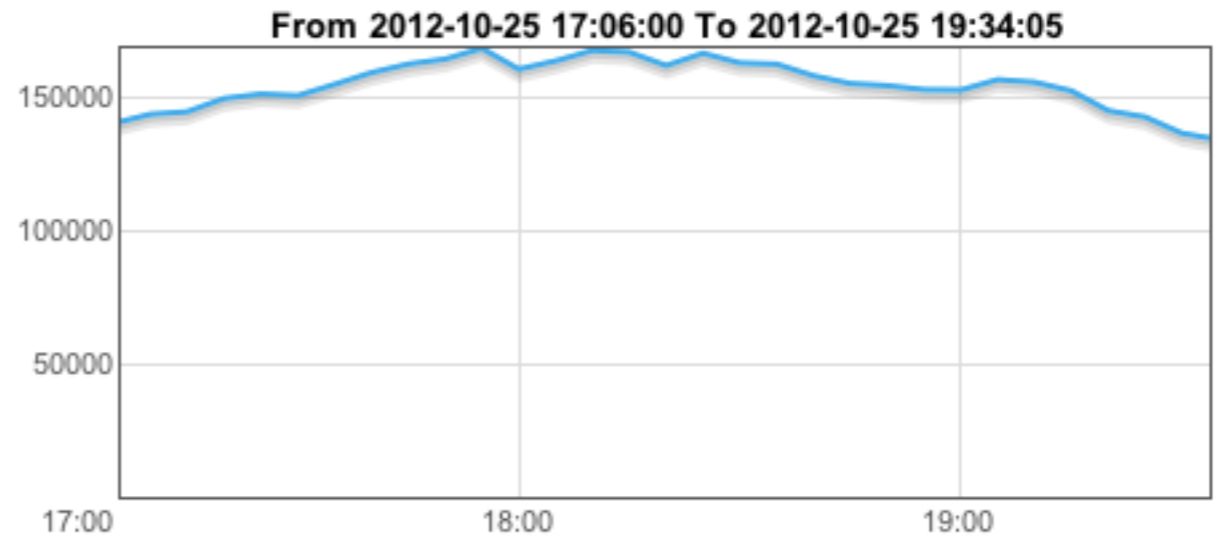
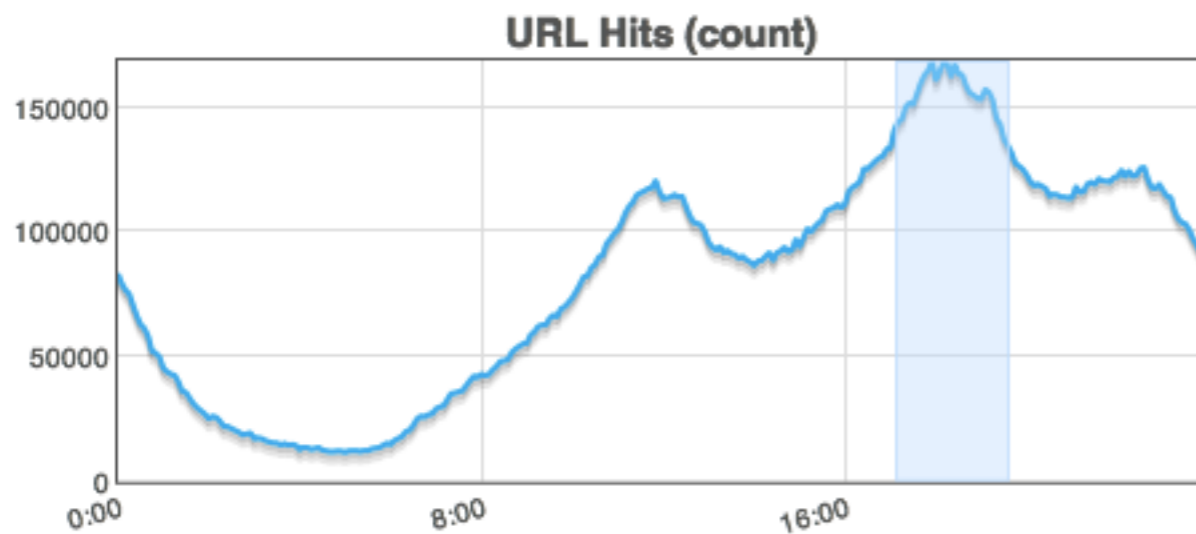
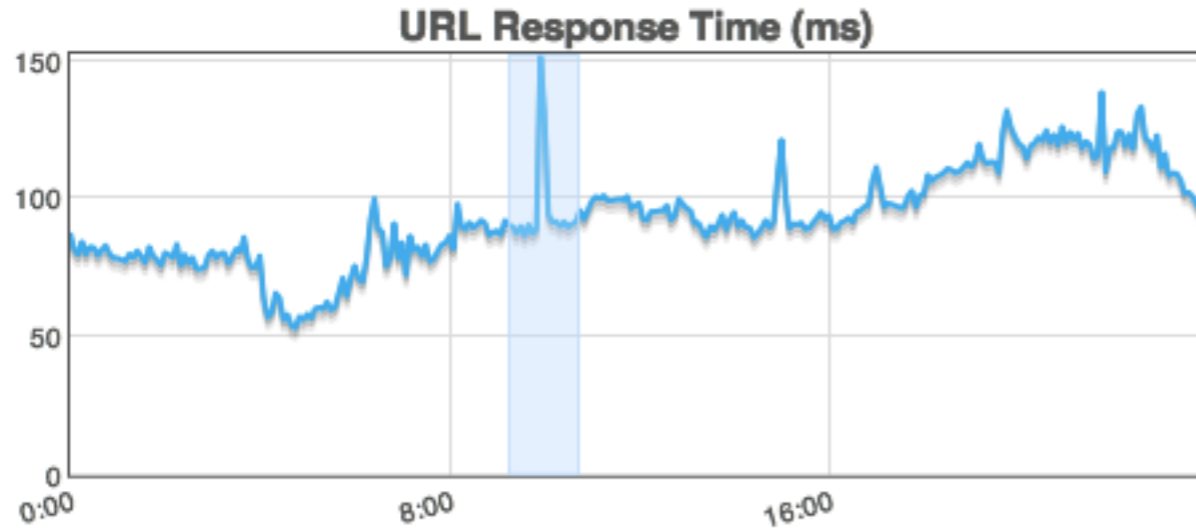
group  Filter 支持多个字符串查询, 例如sql|url|task, 查询结果为包含任一sql、url、task的列

Name	Total Count	Failure Count	Failure%	Sample Link	Min(ms)	Max(ms)	Avg(ms)	95Line(ms)	Std(ms)	TPS	Percent%
TOTAL	977,036	0	0.00%	<a href="#">Log View</a>	1.6	5143.7	198.8	0.0	134.9	0.0	100.00%
<a href="#">[:: show ::] /group/group</a>	48	0	0.00%	<a href="#">Log View</a>	96.2	5046	801.6	2955.8	1111.3	0.0	0.00%
<a href="#">[:: show ::] /note/group</a>	6,973	0	0.00%	<a href="#">Log View</a>	70.3	5056	282.1	574.4	138.2	0.1	0.71%
<a href="#">[:: show ::] /note/group</a>	709,125	0	0.00%	<a href="#">Log View</a>	8.7	5143.7	233.1	455.3	135.1	8.2	72.58%
<a href="#">[:: show ::] /member/group</a>	798	0	0.00%	<a href="#">Log View</a>	70.2	3561.3	193.9	384.3	169.7	0.0	0.08%
<a href="#">[:: show ::] /event/group</a>	1,853	0	0.00%	<a href="#">Log View</a>	16.9	1513.6	158.6	241.9	51.7	0.0	0.19%
<a href="#">[:: show ::] /member/group</a>	97	0	0.00%	<a href="#">Log View</a>	76.9	311.5	140.9	186.8	50.2	0.0	0.01%
<a href="#">[:: show ::] /member/group</a>	1,065	0	0.00%	<a href="#">Log View</a>	47.8	373	109.2	163.2	35.2	0.0	0.11%
<a href="#">[:: show ::] /index/group</a>	255,006	0	0.00%	<a href="#">Log View</a>	49.2	4026.2	102.9	152.7	65.9	3.0	26.10%
<a href="#">[:: show ::] /ajax/joinGro</a>	874	0	0.00%	<a href="#">Log View</a>	2.2	5016	101.7	692.9	505.1	0.0	0.09%
<a href="#">[:: show ::] /group/group</a>	49	0	0.00%	<a href="#">Log View</a>	30.4	114.3	67.3	77.9	24.9	0.0	0.01%
<a href="#">[:: show ::] /group/group</a>	545	0	0.00%	<a href="#">Log View</a>	38.4	382.9	60.9	98.2	27.1	0.0	0.06%
<a href="#">[:: show ::] /group/group</a>	245	0	0.00%	<a href="#">Log View</a>	15.8	100.3	44.1	62.9	14.9	0.0	0.03%
<a href="#">[:: show ::] /s/css/img/g</a>	358	0	0.00%	<a href="#">Log View</a>	1.6	10.8	3.2	4.4	1.2	0.0	0.04%

# Transaction 报表

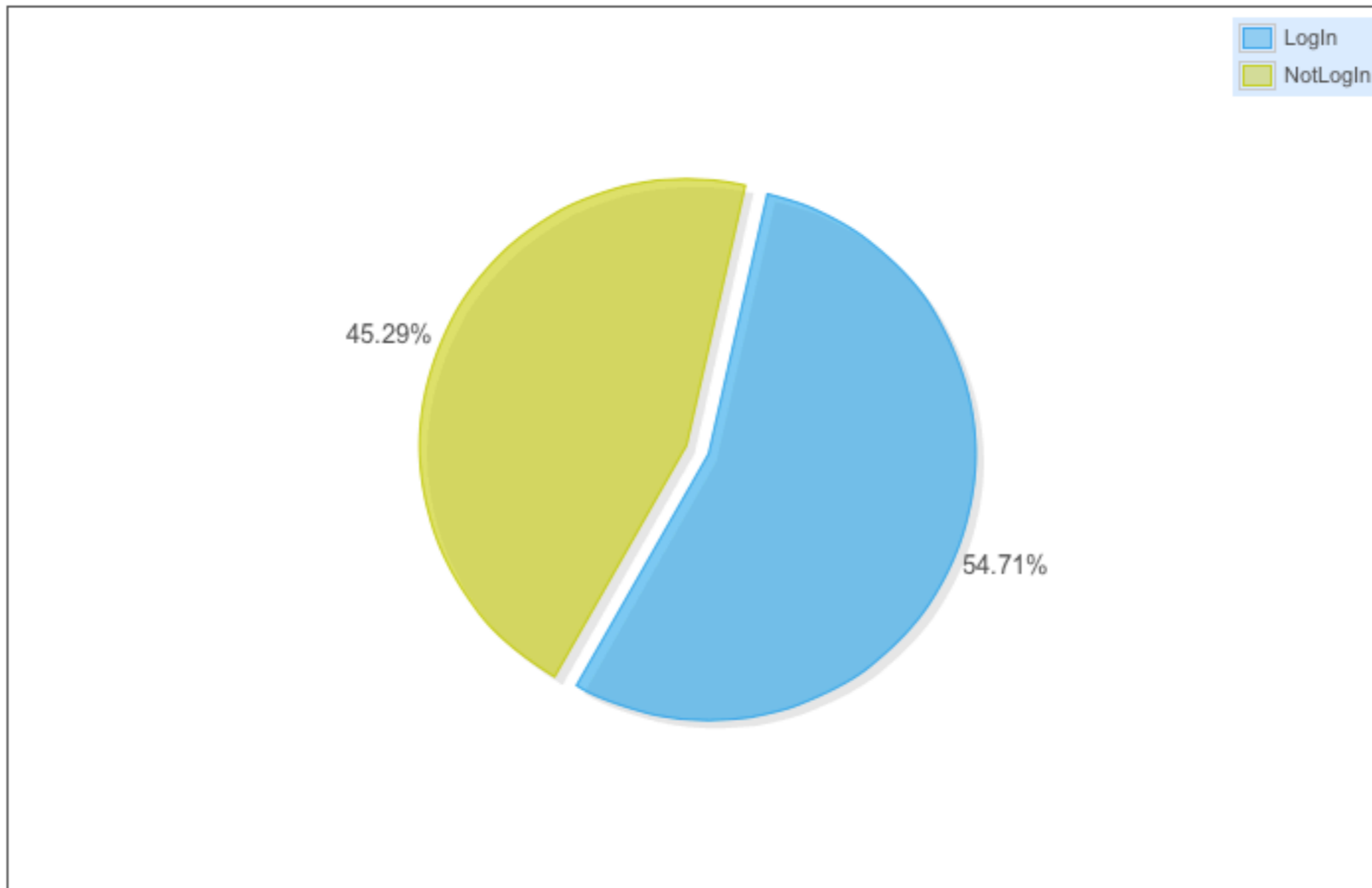


# Transaction 报表



# Event报表

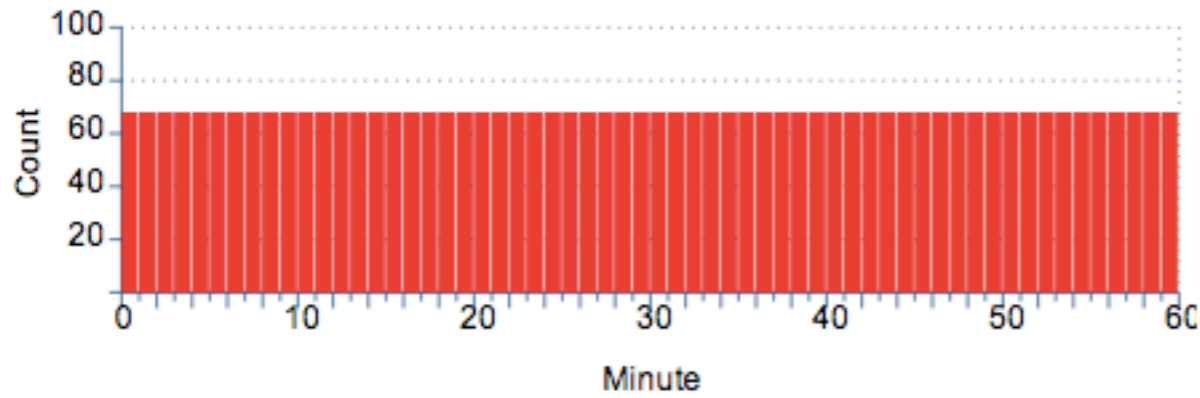
Name	Total Count	Failure Count	Failure%	Sample Link	TPS	Percent%
<a href="#">[:: show ::]</a> NotLogin	9,813,116	0	0.00%	<a href="#">Log View</a>	113.6	45.29%
<a href="#">[:: show ::]</a> Login	11,852,306	0	0.00%	<a href="#">Log View</a>	137.2	54.71%



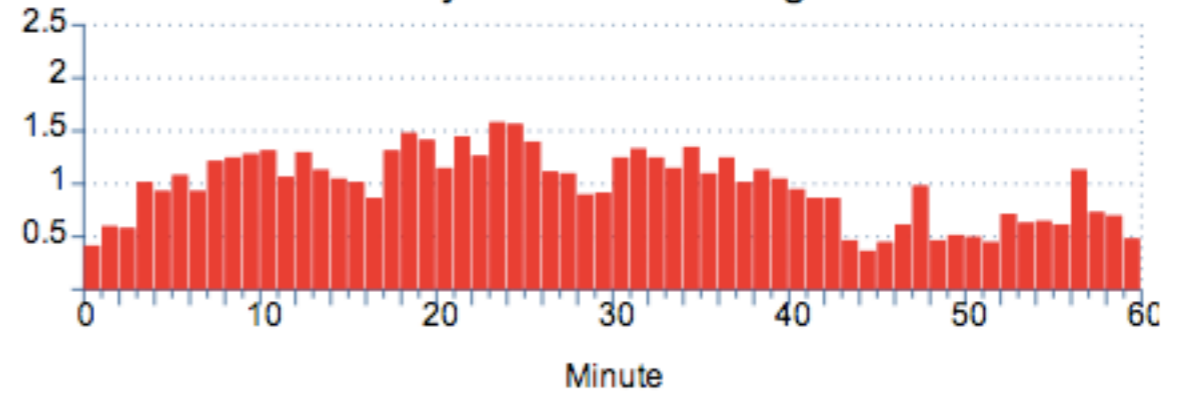


# 心跳报表

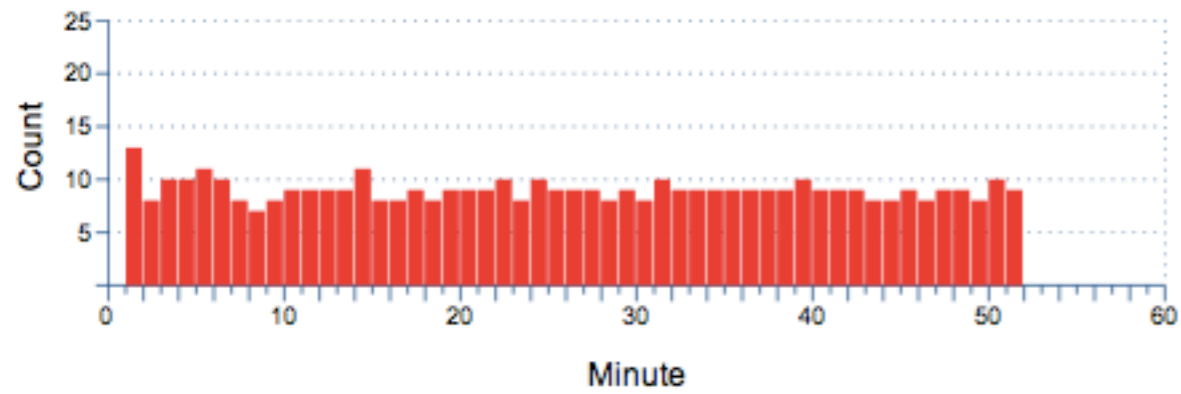
HTTP Thread



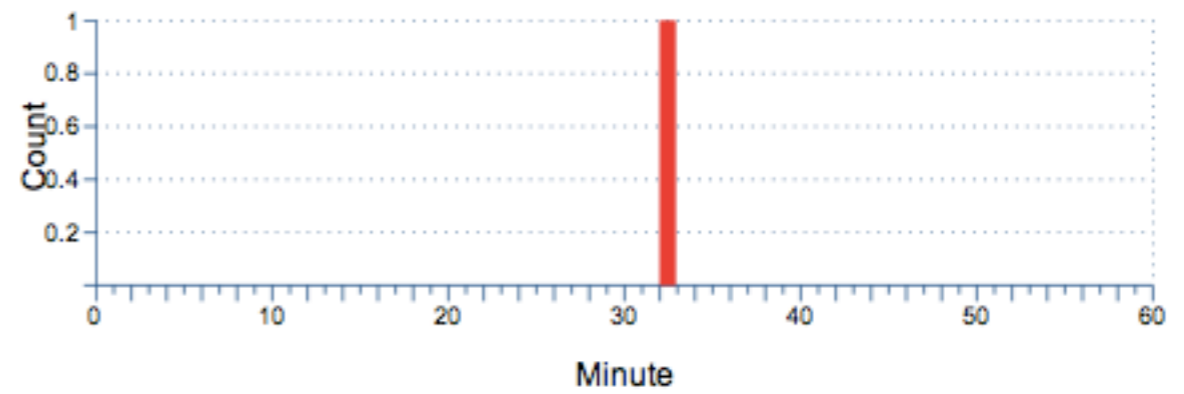
System Load Average



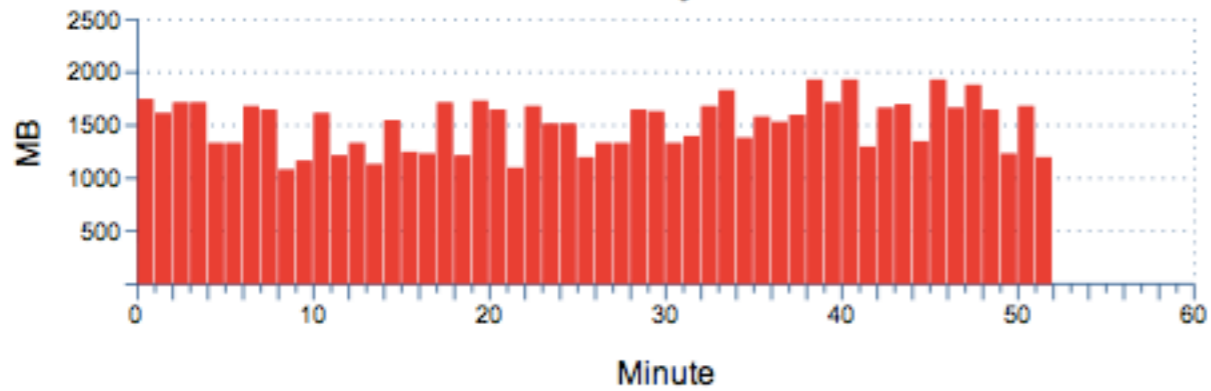
NewGc Count



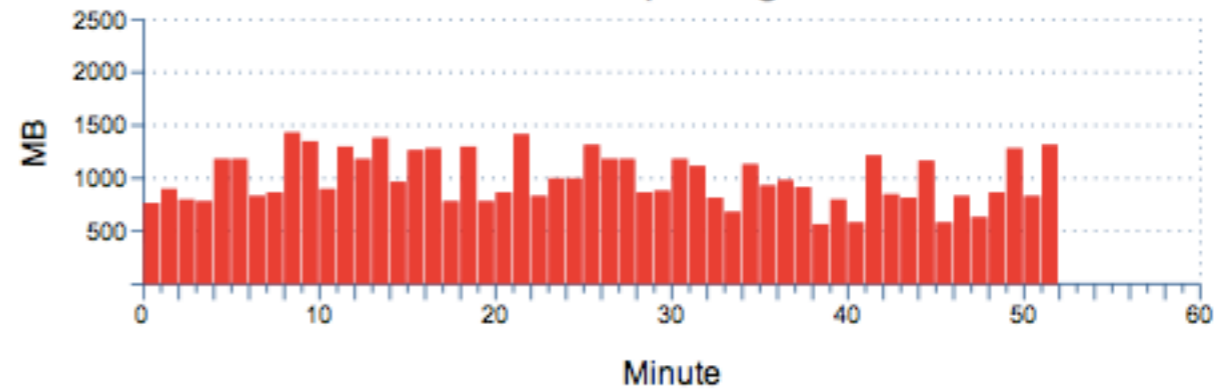
OldGc Count



Memory Free



Heap Usage





# 更多报表

- Matrix

Type	Name	Total Hits	Avg Duration(ms)	Sample Link	Call Ratio			Call Cost		SQL Ratio			SQL Cost	
					Min	Max	Avg	Time(ms)	Time%	Min	Max	Avg	Time(ms)	Time%
URL	/scoreboard.bin	55	195.8	<a href="#">Log View</a>	1	100	65.7	2	80.3%	1	9	1.4	8	06.3%
URL	/hc	12	927.7	<a href="#">Log View</a>	1	172	53.5	14	86.5%	18	124	37.1	0	03.2%
URL	/re	8384	163.5	<a href="#">Log View</a>	1	42	21.9	3	50.7%	0	68	3.1	0	01.4%
URL	/us	524	200.6	<a href="#">Log View</a>	1	61	20.7	4	46.5%	10	72	3.3	0	01.6%
URL	/st	8829	145.6	<a href="#">Log View</a>	1	93	20.6	5	81.7%	14	54	3.8	0	01.9%
URL	/fa	192	122.6	<a href="#">Log View</a>	5	165	19.0	4	72.0%	6	120	11.1	0	06.3%
URL	/us	367	166.3	<a href="#">Log View</a>	1	57	18.4	5	65.1%	0	76	1.5	1	01.0%
URL	/re	112529	137.6	<a href="#">Log View</a>	1	46	18.4	3	50.9%	0	95	2.7	0	01.4%

解读：URL(/scoreboard.bin)平均每次调用了65.7次服务和1.4次数据库访问，服务调用平均耗时约2ms，数据库访问平均耗时8ms，服务调用平均占用了整个URL处理时间的80.3%，数据库访问占6.3%。

建议：优化服务调用，减少服务调用的次数。

# 更多报表

- Cross (客户端)

Type	RemoteProject	Total	Failure	Failure%	Avg(ms)
PigeonCall	AllServers	202,167	2	0.00%	11.40
PigeonCall	FeedServer	78,647	0	0.00%	5.45
PigeonCall	UserBaseService	58,802	0	0.00%	3.39
PigeonCall	UnknownProject	36,625	2	0.01%	15.48
PigeonCall	PayEngine	28,093	0	0.00%	39.50

**解读：**在统计时间段内，调用了约202K次服务，其中调用了FeedServer业务78K次，UserBaseService业务58K次等等。

# 更多报表

- Cross (服务端)

Type	RemoteProject	Total	Failure	Failure%	Avg(ms)
PigeonService	AllClients	53,169,304	0	0.00%	1.34
PigeonService	GroupService	13,723,783	0	0.00%	1.63
PigeonService	ShopWeb	11,680,105	0	0.00%	1.83
PigeonService	GroupWeb	6,266,357	0	0.00%	0.36
PigeonService	UserWeb	5,556,574	0	0.00%	1.28
PigeonService	ShoppicWeb	4,600,250	0	0.00%	1.11
PigeonService	UserBaseService	4,158,036	0	0.00%	1.81
PigeonService	TuanGouWeb	3,779,670	0	0.00%	0.45
PigeonService	ShopSearchWeb	1,800,354	0	0.00%	1.06
PigeonService	PromoWeb	464,401	0	0.00%	0.88
PigeonService	DPIndexWeb	357,403	0	0.00%	1.03
PigeonService	MessageWeb	328,265	0	0.00%	1.44
PigeonService	AuditbackService	255,147	0	0.00%	0.91
PigeonService	PayOrder	88,711	0	0.00%	1.38
PigeonService	AccountWeb	39,480	0	0.00%	1.96
PigeonService	UserService	34,401	0	0.00%	3.47
PigeonService	MobileMembercardMainServer	27,125	0	0.00%	3.57

解读：在统计时间段内，所有服务被调用了约53M次，其中13M次调用来自于GroupService，11M次调用来自于ShopWeb等等。

# 更多报表

- Cache

Type	Total	Missed	Hit Rate(%)	Avg(ms)	TPS
Cache.memcached	53,597,387	2808568	94.76%	0.9	620.3
Cache.web	129,676,145	0	100.00%	0.0	1500.9

Name	Total	Missed	Hit Rate(%)	Avg(ms)	TPS
ALL	26,919,323	2808568	89.57%	0.7	311.6
eUserAtUC:get	23,976,767	1097580	95.42%	0.7	277.5
oUserProfile:get	2,878,562	1691191	41.25%	0.6	33.3
lUserLocation:get	63,994	19797	69.06%	0.6	0.7

# 更多报表

- SQL

Databases: All ] [ [ [ [ ] ] ]

	Table	Total	Failure	Failure%	Avg(ms)	Percent%	TPS
<a href="#">[:: hide ::]</a>	All	99,603,600	0	0.00%	3.08	100.00%	1152.82
Method	Total	Failure	Failure%	Avg(ms)	Percent%	TPS	
Select	99,066,369	0	0.00%	3.09	99.46%	1146.60	
Insert	197,642	0	0.00%	1.84	0.20%	2.29	
Update	321,639	0	0.00%	2.04	0.32%	3.72	
Delete	17,042	0	0.00%	1.39	0.02%	0.20	
Execute	908	0	0.00%	389.60	0.00%	0.01	
<a href="#">[:: show ::]</a>	DP_Group	33,721,213	0	0.00%	0.74	33.86%	390.29
<a href="#">[:: show ::]</a>	DP_GroupNoteExtInfo	22,240,343	0	0.00%	0.72	22.33%	257.41
<a href="#">[:: show ::]</a>	DP_GroupUser	9,004,590	0	0.00%	1.25	9.04%	104.22
<a href="#">[:: show ::]</a>	DP_GroupNote	8,780,891	0	0.00%	14.92	8.82%	101.63
<a href="#">[:: show ::]</a>	GP_GroupSetDetail	6,339,051	0	0.00%	0.53	6.36%	73.37

# 更多报表

- Database

Domains: [ All ] [ GroupService ] [ GroupWeb ]

	Table	Total	Failure	Failure%	Avg(ms)	Percent%	TPS
[:: hide ::]	All	99,603,595	0	0.00%	3.08	100.00%	1152.82
Method	Total	Failure	Failure%	Avg(ms)	Percent%	TPS	
Select	99,066,364	0	0.00%	3.09	99.46%	1146.60	
Insert	197,642	0	0.00%	1.84	0.20%	2.29	
Update	321,639	0	0.00%	2.04	0.32%	3.72	
Delete	17,042	0	0.00%	1.39	0.02%	0.20	
Execute	908	0	0.00%	389.60	0.00%	0.01	
[:: show ::]	DP_Group	33,721,210	0	0.00%	0.74	33.86%	390.29
[:: show ::]	DP_GroupNoteExtInfo	22,240,343	0	0.00%	0.72	22.33%	257.41
[:: show ::]	DP_GroupUser	9,004,589	0	0.00%	1.25	9.04%	104.22
[:: show ::]	DP_GroupNote	8,780,891	0	0.00%	14.92	8.82%	101.63
[:: show ::]	GP_GroupSetDetail	6,339,051	0	0.00%	0.53	6.36%	73.37
[:: show ::]	DP_GroupFollowNote	5,828,541	0	0.00%	13.94	5.85%	67.46

解读：此数据库有较高的SELECT请求，平均每秒1146次查询。



# Agenda

- 介绍
- 报表
- 设计
- 未来

# 设计

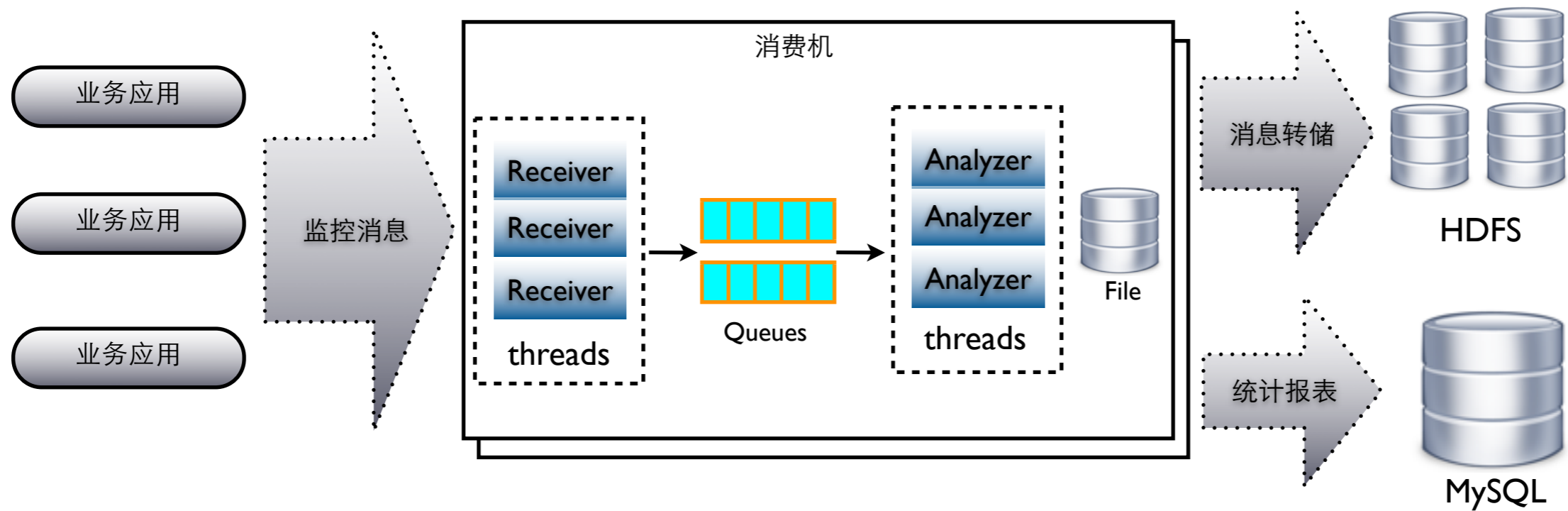
- API
- 实时分析
- 数据建模
- 存储

# API

```
① Transaction t = Cat.newTransaction("URL", pageName);  
  
try {  
② Cat.logEvent("URL.Server", req.getServerName(), SUCCESS, "RemoteIP=" + req.getRemoteAddr() + "&...");  
② Cat.logEvent("URL.Method", req.getScheme() + "/" + req.getMethod(), SUCCESS, req.getRequestURI());  
  
③ processInbound();  
③ processTransition();  
③ processOutboundbound();  
  
④ t.setStatus(SUCCESS);  
} catch (Exception e) {  
④ t.setStatus(e);  
} finally {  
⑤ t.complete();  
}
```

- ① 创建Transaction
- ② 记录子Event
- ③ 记录子Transaction
- ④ 设置状态
- ⑤ 结束Transaction

# 实时处理



# 数据建模

- 目标模型定义
- 访问, 转换和合并
- 模型持久化
  - XML, JSON, Binary, ...
- 代码生成

```
<?xml version="1.0" encoding="UTF-8"?>
<model>
  <entity name="transaction-report" root="true">
    <attribute name="domain" value-type="String" key="true" />
    <attribute name="startTime" value-type="Date" />
    <attribute name="endTime" value-type="Date" />
    <entity-ref name="machine" type="map" names="machines" />
  </entity>
  <entity name="machine">
    <attribute name="ip" value-type="String" key="true"/>
    <entity-ref name="type" type="map" names="types" />
  </entity>
  <entity name="type">
    <attribute name="id" value-type="String" key="true" />
    <attribute name="total-count" value-type="int" />
    <attribute name="fail-count" value-type="int" />
    <attribute name="min" value-type="double" />
    <attribute name="max" value-type="double" />
    <attribute name="sum" value-type="double" />
    <attribute name="sum2" value-type="double" />
    <element name="success-message" value-type="String" />
    <element name="fail-message" value-type="String" />
    <entity-ref name="name" type="map" names="names" />
  </entity>
  . . .
</model>
```

```
public interface IVisitor {
  public void visitTransactionReport(TransactionReport transactionReport);
  public void visitMachine(Machine machine);
  public void visitType(TransactionType type);
  public void visitName(TransactionName name);
  public void visitRange(Range range);
  public void visitDuration(Duration duration);
}
```

# 存储需求

- 当前小时：读写；历史消息：只读
- 数量：几百万/小时
- 大小：1~100 KB/消息
- 消息ID：生成和使用可能不是同一应用
- 随机读取：消息ID -> Logview
- 顺序读取：Hadoop M/R, Hive -> HDFS
- 压缩
- 归档



# 存储设计

- 介质
  - 内存：容量小, 成本高, 速度快, 顺序和随机读写
  - 本地磁盘：容量有限, 成本低, 本地访问, 顺序和随机读写
  - HDFS：容量大, 成本低, 冗余备份, 顺序读写随机读
  - MySQL：开销大, 成本高, 顺序和随机读写
- 存储
  - 最近2小时
    - 报表：内存
    - 消息：本地磁盘
  - 历史小时
    - 报表：MySQL（历史日报表，周报表，月报表以小时报表为基础生成）
    - 消息：HDFS

# 消息存储设计

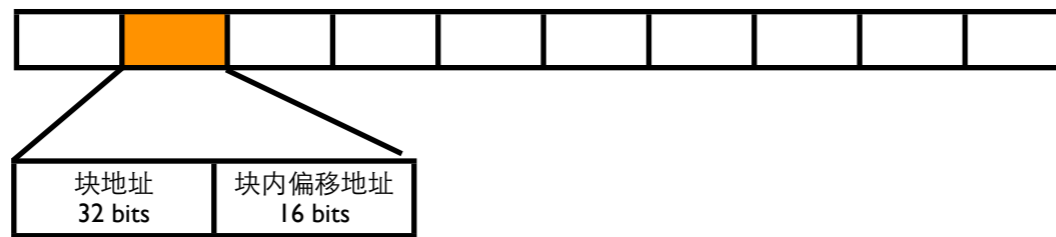
消息ID: **ShopWeb-0a010680-375030-2**

/2012/10/13/14/ShopService-ShopWeb-10.1.6.1  
/2012/10/13/14/ShopService-ShopWeb-10.1.6.2

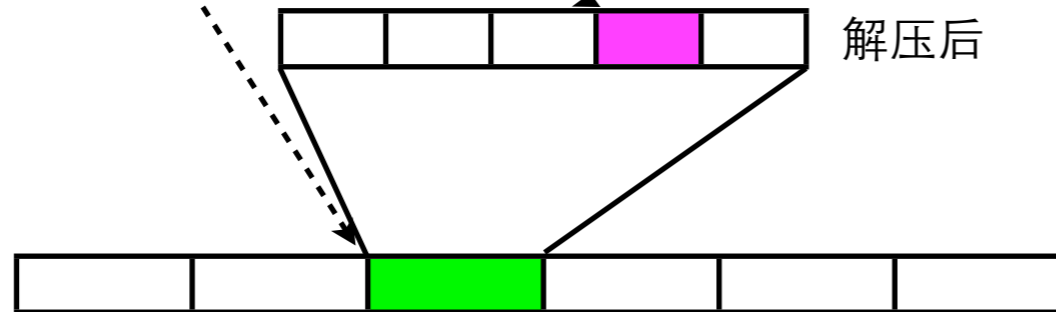
375030 => 2012-10-13 14:00:00  
ShopService => 消息被记录的domain  
10.1.6.1/2 => 消息被处理的机器IP  
0a010680 => 10.1.6.128 用于保证消息ID唯一性

顺序读: /2012/10/13/14/ShopService-\*-\*  
随机读: /2012/10/13/14/\*-ShopWeb-\*

索引



数据



GZIP压缩, 压缩前大小<64K

# 心得

- 先做小做精，再做大做全
- 持续集成，持续发布，不断监控
- 单机开发和调试
- **Everything Fails**
- 关注客户，快速响应

# Agenda

- 介绍
- 报表
- 设计
- 未来

# 未来

- 告警服务
- 数据服务
- 故障定位
- 支持Hive分析
- 前端监控
- 代码开源

谢谢！