利用亚马逊新服务实时洞察数据

——实时数据检索与云时代的即时商业智能

程浩鑫

实时数据检索
Amazon Elasticsearch

您的数据正在不断增长

您有什么选择?

Elasticsearch怎样提供帮助

elasticsearch.

- 一个强大的、实时的、分布式的、开源的搜索和分析引擎
- 建立在Apache Lucene之上
- Apache 2.0 协议
- 面向开发人员的RESTful API

使用管理Elasticsearch却没那么简单

“Elasticsearch 让我们使用 ELK轻松和快速构建前沿大数据分析应用。Amazon Elasticsearch 服务提供直接访问 Elasticsearch API，并且省去了管理任务，给我们提供需要的可管理性、灵活性和控制”

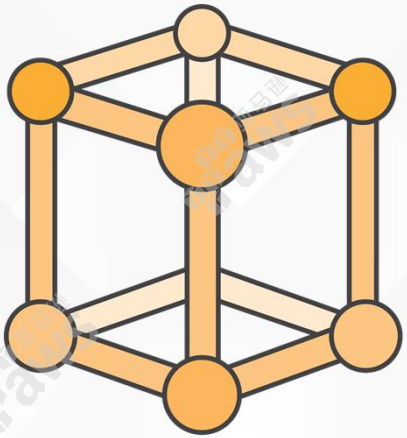
“Elasticsearch allows us to easily and quickly build bleeding edge big data and analytics applications using the ELK stack. By offering direct access to the Elasticsearch API while offloading administrative tasks, Amazon Elasticsearch Service gives us the manageability, flexibility and control we need ”

Sean Curtis,
SVP Engineering at Major League Baseball Advanced Engineering

什么是 Amazon Elasticsearch 服务

- Amazon Elasticsearch 服务托管于 AWS， 使得它易于在云端构建，操作和扩展Elasticsearch 集群

Amazon Elasticsearch 核心价值



简单的集群创建
和配置管理

elasticsearch.

kibana

ELK支持

AWS IAM安全
Amazon
CloudWatch监控
AWS CloudTrail审计

与其他AWS服务集
成选项(CloudWatch
Logs, Amazon
DynamoDB,
Amazon S3,
Amazon Kinesis)

创建一个集群

Configure cluster

Configure a cluster based on your traffic, data, and availability requirements. A cluster is a collection of one or more data nodes (instances) that holds your data and provides indexing and search capabilities across all nodes.

Node configuration

If you have a large amount of data to upload or anticipate a large volume of search requests, you can preconfigure your domain with additional resources. Set the instance type and instance count based on the size of Elasticsearch indices, shards, and replicas that you intend to create on your cluster.

Instance type m3.medium.elasticsearch (def…

Instance count 1 (default)

Enable dedicated master

Enable zone awareness

Storage configuration

Choose a storage type for your data nodes. Storage types do not apply to dedicated master nodes.

Storage type Instance (default)

Snapshot configuration

Once a day, Amazon ES takes an automated snapshot of your cluster. You can set the start hour for the snapshot. We recommend that you choose a time when traffic on your cluster is low.

Automated snapshot start hour 00:00 UTC (default)

实例类型建议

实例	适用场景
T2	入门级，开发和测试。
M3	读写负载相同，最大支持5TB的EBS存储。
R3	大量读取操作或很多查询工作负载 (例如，聚合查询)。
I2	存储优化实例，最大支持16TB SSD实例存储。

AWS CLI 命令

add-tags
create-elasticsearch-domain
delete-elasticsearch-domain
describe-elasticsearch-domain
describe-elasticsearch-domain-config
describe-elasticsearch-domains
list-domain-names
list-tags
remove-tags
update-elasticsearch-domain-config

```
aws es create-elasticsearch-domain --domain-name my-domain
 --elasticsearch-cluster-config
   InstanceType=m3.xlarge.elasticsearch,InstanceCount=3
 --ebs-options
   EBSEnabled=true,VolumeType=gp2,VolumeSize=512
```

Amazon Elasticsearch的域

用于监控的CloudWatch/CloudTrail

单端点, REST API

Elasticsearch API

Amazon Route 53

IAM

IAM 集成

Elastic Load Balancing

节点

安全访问您的域

安全访问您的域

指定的用户，通过签名的请求访问

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "",
      "Effect": "Allow",
      "Principal": {
       "AWS": "arn:aws:iam:123456789012:user/susan"
      },
      "Action": [ "es:ESHttpGet", "es:ESHttpPut", "es:ESHttpPost",
                  "es:CreateElasticsearchDomain",
                  "es:ListDomainNames" ],
      "Resource":
         "arn:aws:es:us-east-1:###:domain/logs-domain/<index>/*"
} ] }
```

安全访问您的域

每个策略允许/拒绝HTTP方法和配置操作

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "",
      "Effect": "Allow",
      "Principal": {
        "AWS": "arn:aws:iam:123456789012:user/susan"
      },
      "Action": [ "es:ESHttpGet", "es:ESHttpPut", "es:ESHttpPost",
                  "es:CreateElasticsearchDomain",
                  "es:ListDomainNames" ],
      "Resource":
         "arn:aws:es:us-east-1:###:domain/logs-domain/<index>/*"
} ] }
```

安全访问您的域

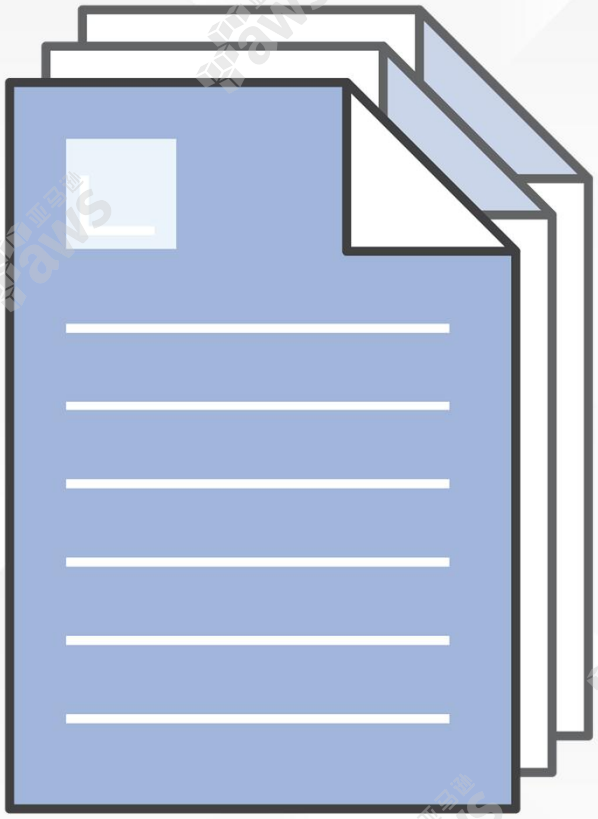
索引级别的粒度控制

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "",
      "Effect": "Allow",
      "Principal": {
        "AWS": "arn:aws:iam:123456789012:user/susan"
      },
      "Action": [ "es:ESHttpGet", "es:ESHttpPut", "es:ESHttpPost",
                  "es:CreateElasticsearchDomain",
                  "es:ListDomainNames" ],
      "Resource":
        "arn:aws:es:us-east-1:###:domain/logs-domain/<index>/*"
} ] }
```

安全访问您的域

And/Or 使用基于IP的访问控制

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "",
      "Effect": "Allow",
      "Principal": {
        "AWS": "*"
      },
      "Action": [ "es:ESHttpGet", "es:ESHttpPut", "es:ESHttpPost",
                  "es:CreateElasticsearchDomain",
                  "es:ListDomainNames" ],
      "Resource":
        "arn:aws:es:us-east-1:###:domain/logs-domain/<index>/*",
      "Condition":
        "IpAddress": {
          "aws:SourceIp": [ "xx.xx.xx.xx/yy" ]
        } } ] }
```

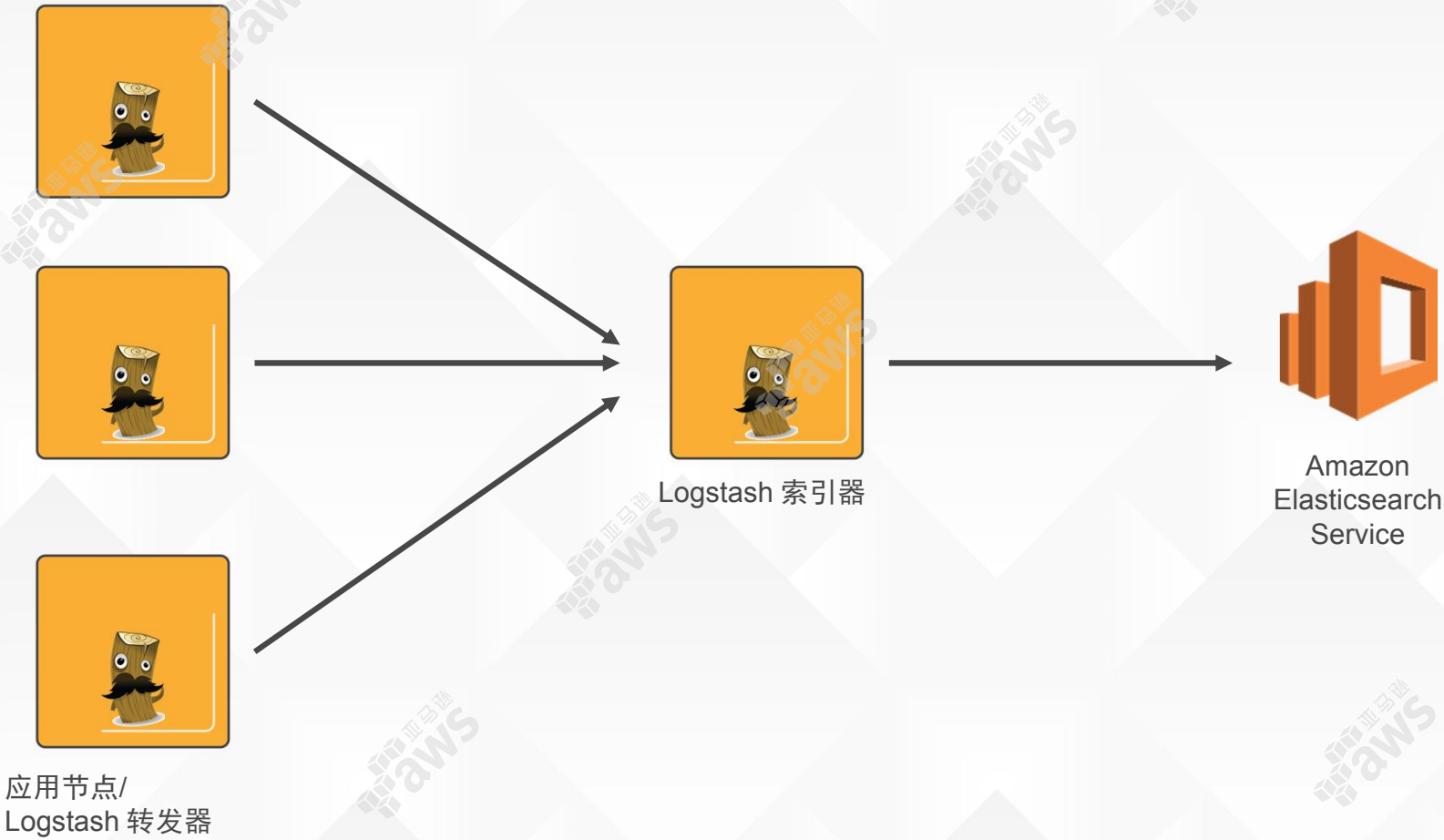


装载数据

直接访问Elasticsearch API

```
$ curl -XPUT https://<endpoint>/blog -d '{
   "settings" : { "number_of_shards" : 3, "number_of_replicas" : 1 } }'
$ curl -XPOST http://<endpoint>/blog/post/1 -d '{
    "author":"jon handler",
    "title":"Amazon ES Launch" }'
$ curl -XPOST https://<endpoint>/blog/post/_bulk -d '
    { "index" : { "_index" : "blog", "_type" : "post", "_id" : "2"}}
    {"title":"Amazon ES for search", "author": "pravin pillai"},
    { "index" : { "_index":"blog", "_type":"post", "_id":"3" } }
    { "title":"Analytics too", "author": "vivek sriram"}'
$ curl -XGET http://<endpoint>/_search?q=ES
{"took":16,"timed_out":false,"_shards":{"total":3,"successful":3,"failed":0},"hits":{"t
otal":2,"max_score":0.13424811,"hits":[{"_index":"blog","_type":"post","_id":"1","_scor
e":0.13424811,"_source":{"author":"jon handler", "title":"Amazon ES Launch"
}},{"_index":"blog","_type":"post","_id":"2","_score":0.11506981,"_source":{"title":"Am
azon ES for search", "author": "pravin pillai"},}]}}
```

使用Logstash读取数据

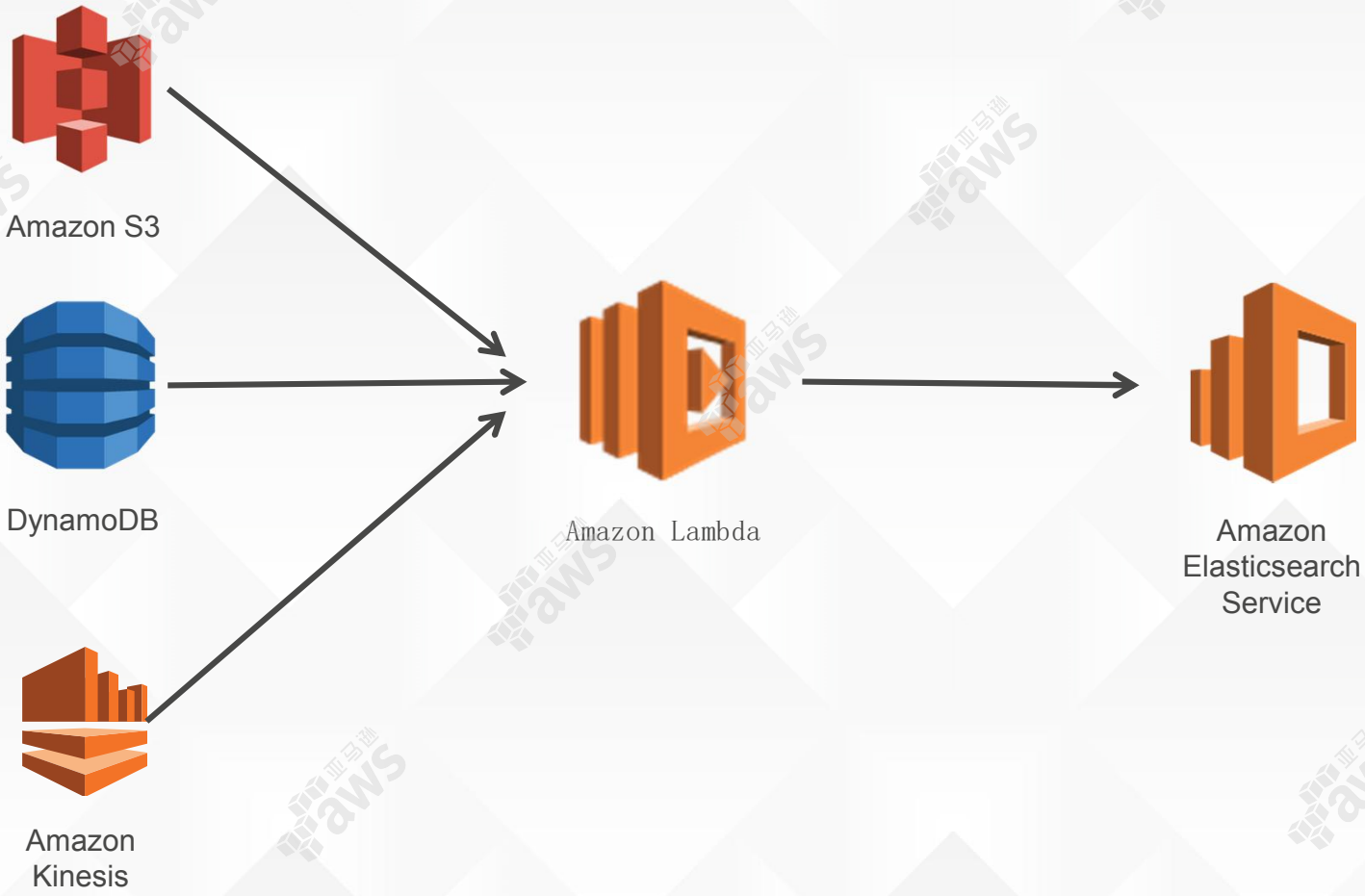


Amazon ES Logstash 插件

https://github.com/awslabs/logstash-output-amazon_es

```
output {
  amazones {
    *hosts => ["foo.us-east-1.es.amazonaws.com"]
    *region => "us-east-1"
    access_key => 'ACCESS_KEY' (optional)
    secret_key => 'SECRET_KEY' (optional)
    index => "logstash-%{+YYYY.MM.dd}"
  }
}
```

使用Lambda读取数据

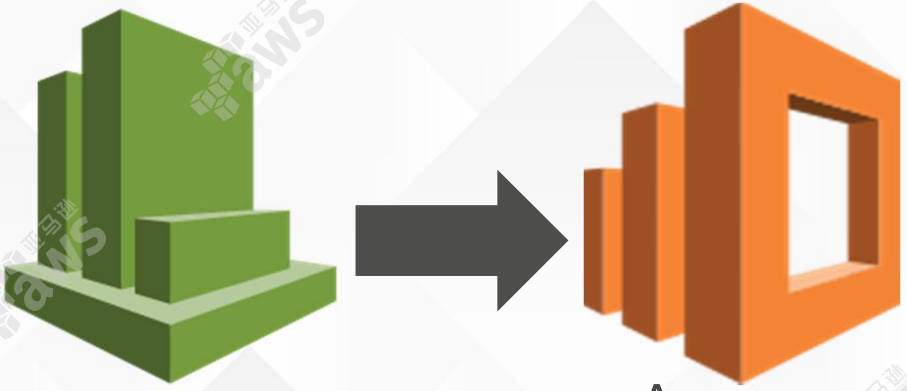


Lambda上传代码片段 (node.js)

```
var AWS = require('aws-sdk');
var creds = new AWS.EnvironmentCredentials('AWS');

function postDocumentToES(doc, context) {
    var req = new AWS.HttpRequest(endpoint);
    var signer = new AWS.Signers.V4(req, 'es');
    signer.addAuthorization(creds, new Date());
    var send = new AWS.NodeHttpClient();
    send.handleRequest(req, null, function(httpResp)...
```

https://github.com/awslabs/amazon-elasticsearch-lambda-samples



导出日志到 Amazon ES

数据持久性 - 快照和恢复

每日自动快照

- 没有附加费用
- 快照保存14天

手动快照

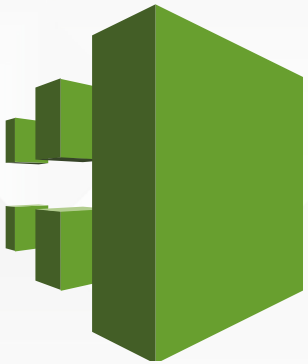
- 注册S3存储桶

```
curl -XPUT http://<endpoint>/_snapshot/<repo-name>
  -d '{"type":"s3",
       "settings": {
         "bucket":"<bucket>",
         "region":"<region>",
         "role-arn":"<arn>"}}'
```

- 快照命令

```
curl -XPUT http://<endpoint>/_snapshot/<repo-name>/snapshot1
```

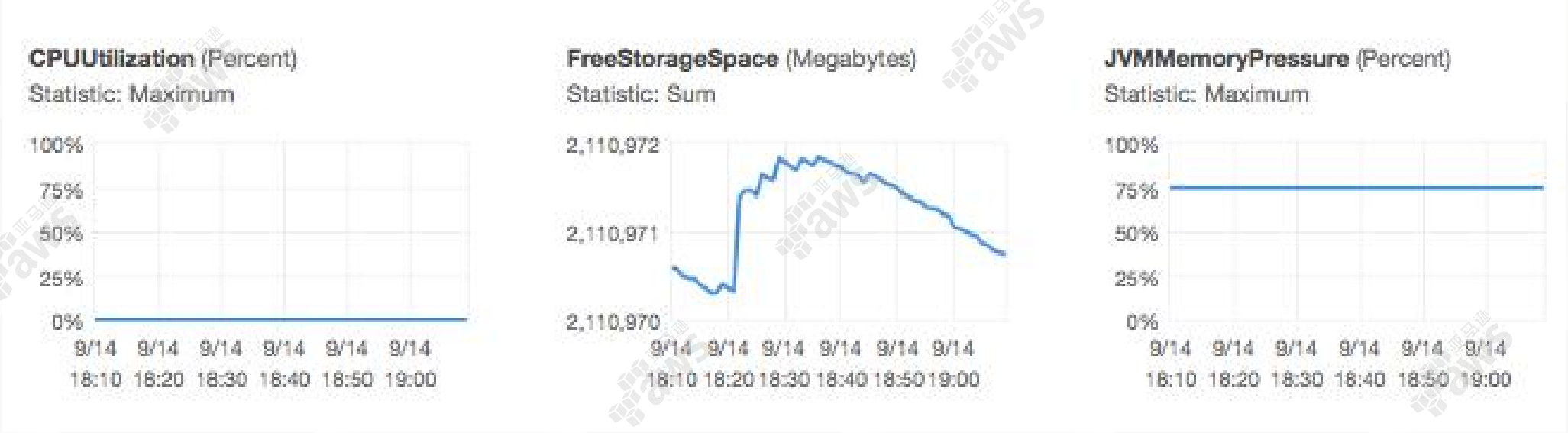
CloudWatch



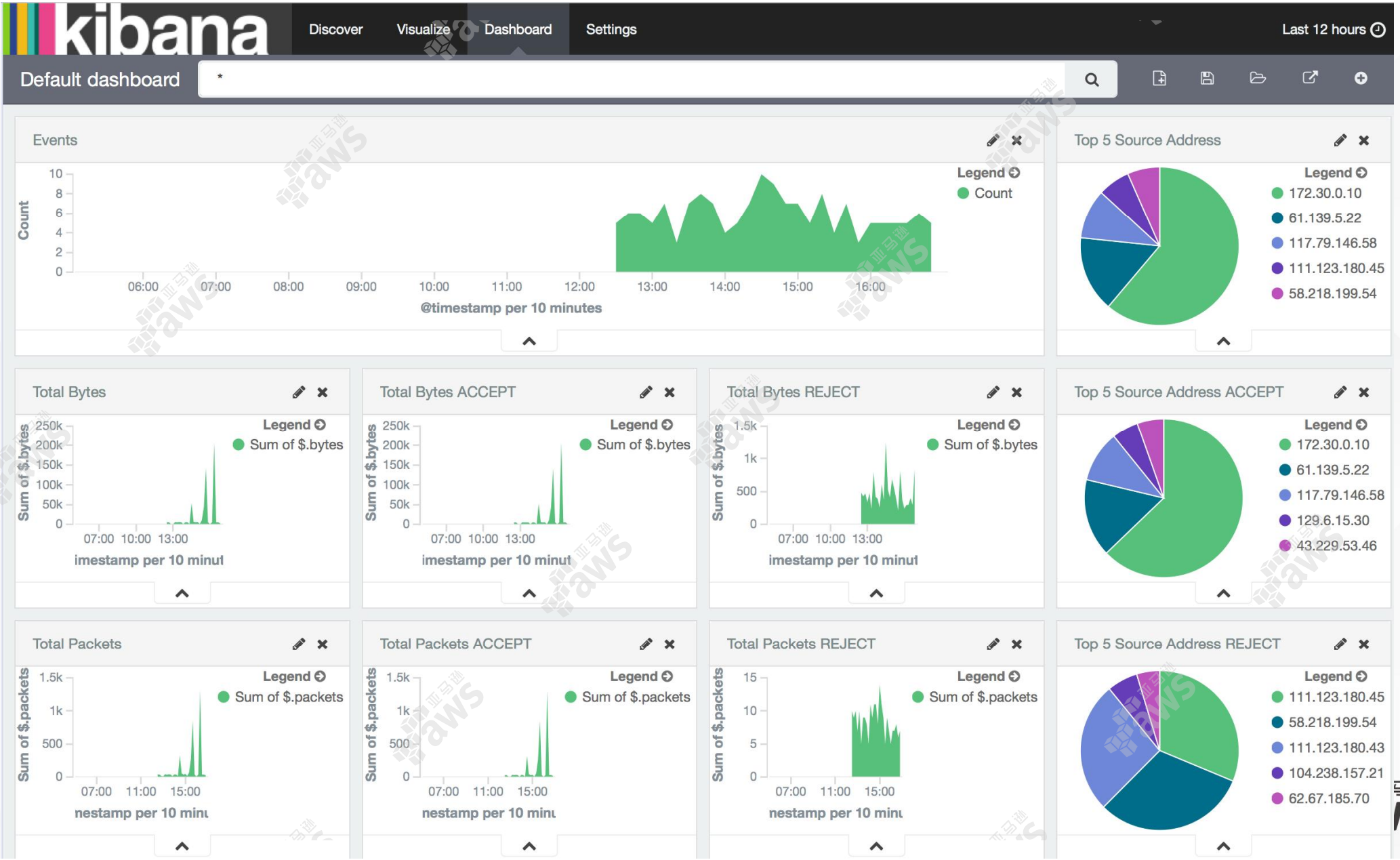
CloudTrail

监控和审计

监控

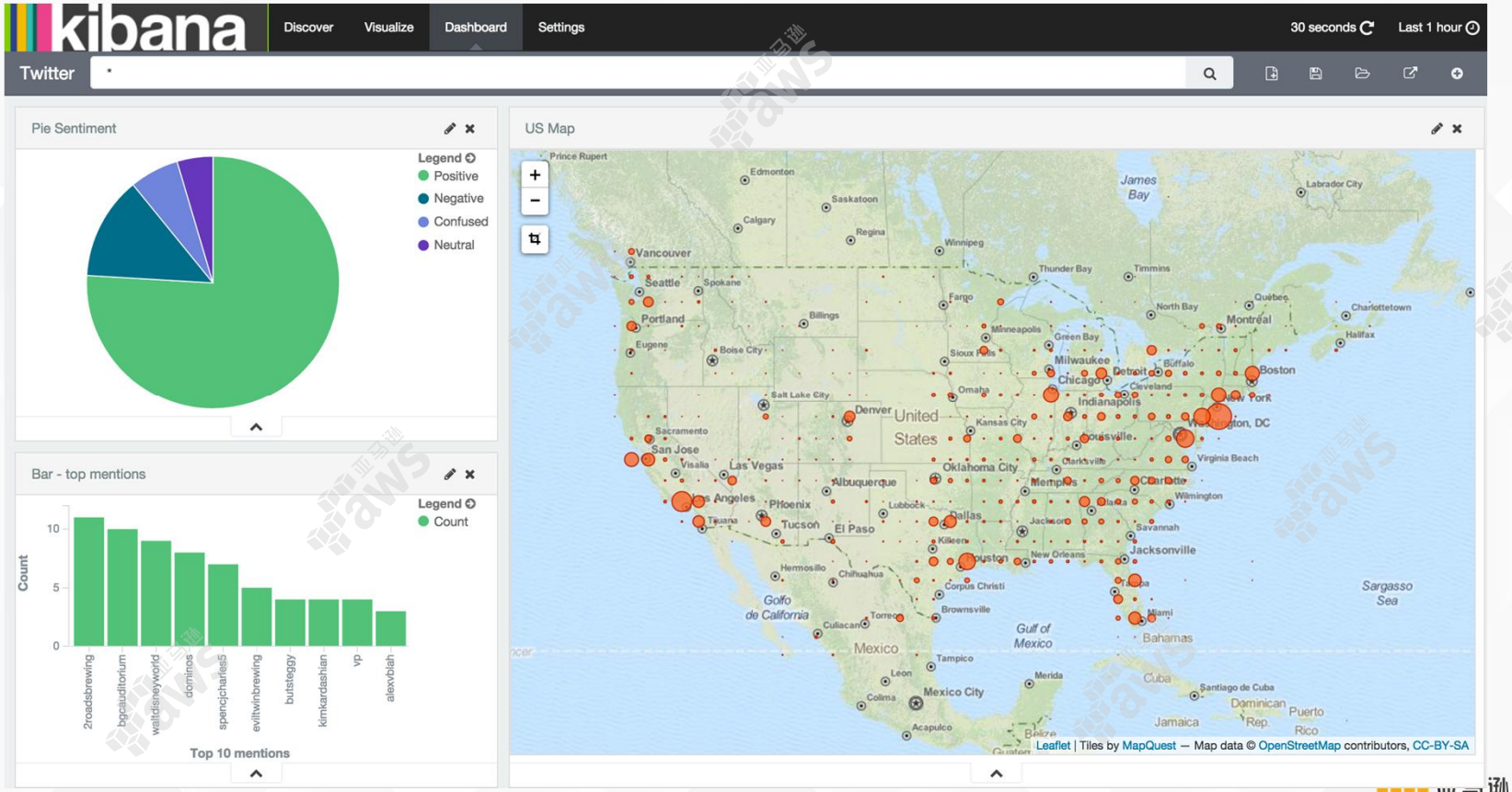

CPUUtilization (Percent)
Statistic: Maximum
100%
75%
50%
25%
0%
9/14 18:10
9/14 18:20
9/14 18:30
9/14 18:40
9/14 18:50
9/14 19:00
FreeStorageSpace (Megabytes)
Statistic: Sum
2,110,972
2,110,971
2,110,970
9/14 18:10
9/14 18:20
9/14 18:30
9/14 18:40
9/14 18:50
9/14 19:00
JVMMemoryPressure (Percent)
Statistic: Maximum
100%
75%
50%
25%
0%
9/14 18:10
9/14 18:20
9/14 18:30
9/14 18:40
9/14 18:50
9/14 19:00

内置Kibana

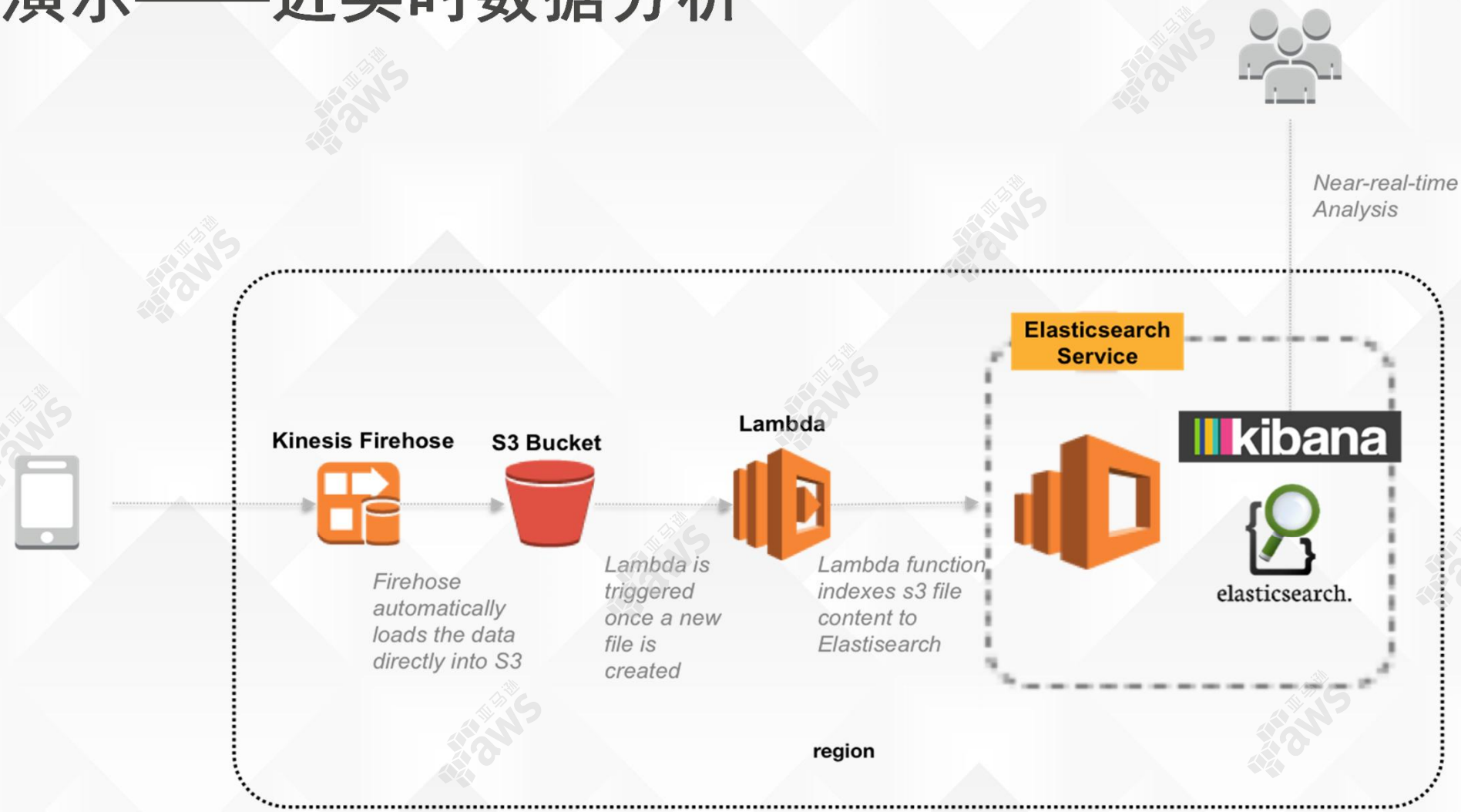

kibana
Discover
Visualize
Dashboard
Settings
Last 12 hours
Default dashboard
*
Events
Count
@timestamp per 10 minutes
Legend
Count
Top 5 Source Address
Legend
172.30.0.10
61.139.5.22
117.79.146.58
111.123.180.45
58.218.199.54
Total Bytes
Sum of $.bytes
Legend
Sum of $.bytes
Total Bytes ACCEPT
Sum of $.bytes
Legend
Sum of $.bytes
Total Bytes REJECT
Sum of $.bytes
Legend
Sum of $.bytes
Top 5 Source Address ACCEPT
Legend
172.30.0.10
61.139.5.22
117.79.146.58
129.6.15.30
43.229.53.46
Total Packets
Sum of $.packets
Legend
Sum of $.packets
Total Packets ACCEPT
Sum of $.packets
Legend
Sum of $.packets
Total Packets REJECT
Sum of $.packets
Legend
Sum of $.packets
Top 5 Source Address REJECT
Legend
111.123.180.45
58.218.199.54
111.123.180.43
104.238.157.21
62.67.185.70

演示——近实时数据分析

- 分析twitter的数据
  - 情绪
  - 引用
  - 分布信息



演示——近实时数据分析


Near-real-time
Analysis
Elasticsearch
Service
Kinesis Firehose
S3 Bucket
Lambda
kibana
elasticsearch.
Firehose
automatically
loads the data
directly into S3
Lambda is
triggered
once a new
file is
created
Lambda function
indexes s3 file
content to
Elastisearch
region

云时代的即时商业智能
Amazon QuickSight

客户面临的挑战

我的营销活动怎样执行?

哪些设备正处于维护状态?

为什么我最有利的地区没有增长?

大量的数据

大量的问题

很少的办法

谁是我的客户?他们购买了什么?

我有多少库存?

我的员工满意度趋势如何?

我的产品盈利能力的区域是什么?

欺诈行为导致的损失增加了么?

传统的商业智能

成本太高

	Named User Plus	Software Update License & Support	Processor License	Software Update License & Support
Business Intelligence Technology Products				
Business Intelligence				
Standard Edition One	1,200	264.00	-	-
Suite Enterprise Edition Plus	2,000	440.00	221,250	48,675.00
Suite Enterprise Edition Plus Upgrade Only	230	50.60	34,500	7,590.00
Server Enterprise Edition	350	77.00	51,800	11,396.00
Business Intelligence Publisher	460	101.20	46,000	10,120.00
Business Intelligence Foundation Suite	3,675	808.50	300,000	66,000.00
Disconnected Analytics	580	127.60	-	-
Server Administrator	5,800	1,276.00	-	-
Scorecard and Strategy Management	995	218.90	89,550	19,701.00
Business Intelligence Mobile	360	79.20	-	-
Business Intelligence Server Enterprise Edition Options:				
Interactive Dashboard	580	127.60	86,500	19,030.00
Delivers	350	77.00	51,800	11,396.00
Answers	580	127.60	86,500	19,030.00
Office Plug-in	230	50.60	34,500	7,590.00
Reporting and Publishing	460	101.20	70,000	15,400.00
Business Intelligence Suite Enterprise Edition Plus Options:				
Business Intelligence Management Pack	230	50.60	11,500	2,530.00
Business Intelligence Data Integration Technology				
Data Integrator for Oracle Business Intelligence	690	151.80	23,000	5,060.00
Informatica PowerCenter and PowerConnect Adapters	690	151.80	25,300	5,566.00
Metadata Management for Oracle Business Intelligence	2,400	528.00	80,000	17,600.00

在看到第一次分析前就要支付数百万美元

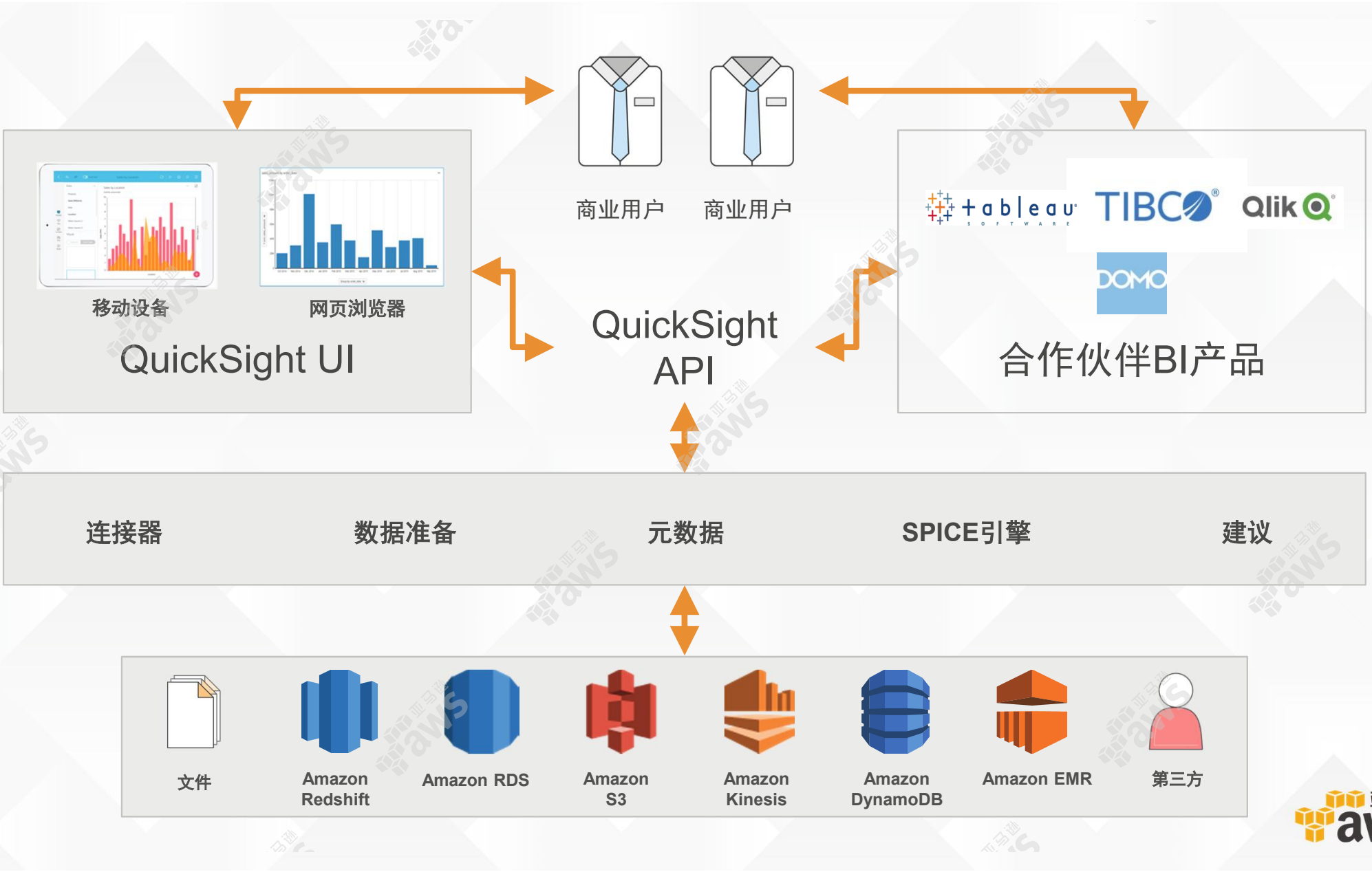
3年TCO 150到250美元每月每用户

花费时间太长

花费6到12个月的咨询和软件实现

什么是Amazon QuickSight?

快速的，云端的，商业智能服务，成本是传统BI软件十分之一


商业用户
商业用户
移动设备
网页浏览器
QuickSight UI
QuickSight API
合作伙伴BI产品
连接器
数据准备
元数据
SPICE引擎
建议
文件
Amazon Redshift
Amazon RDS
Amazon S3
Amazon Kinesis
Amazon DynamoDB
Amazon EMR
第三方

主要特点

- 易于探索AWS数据
- 使用SPICE快速获得业务见解
  - **S**uper-fast, **P**arallel, **I**n-memory, **C**alculation **E**ngine
- 使用AutoGraph进行转换并获得直观的可视化内容
- 利用StoryBoard安全分享并协作

易于探索AWS数据

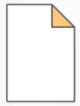

Amazon EMR


Amazon RDS

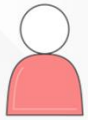

Amazon Kinesis


Amazon S3


Amazon Dynamo DB


文件上传

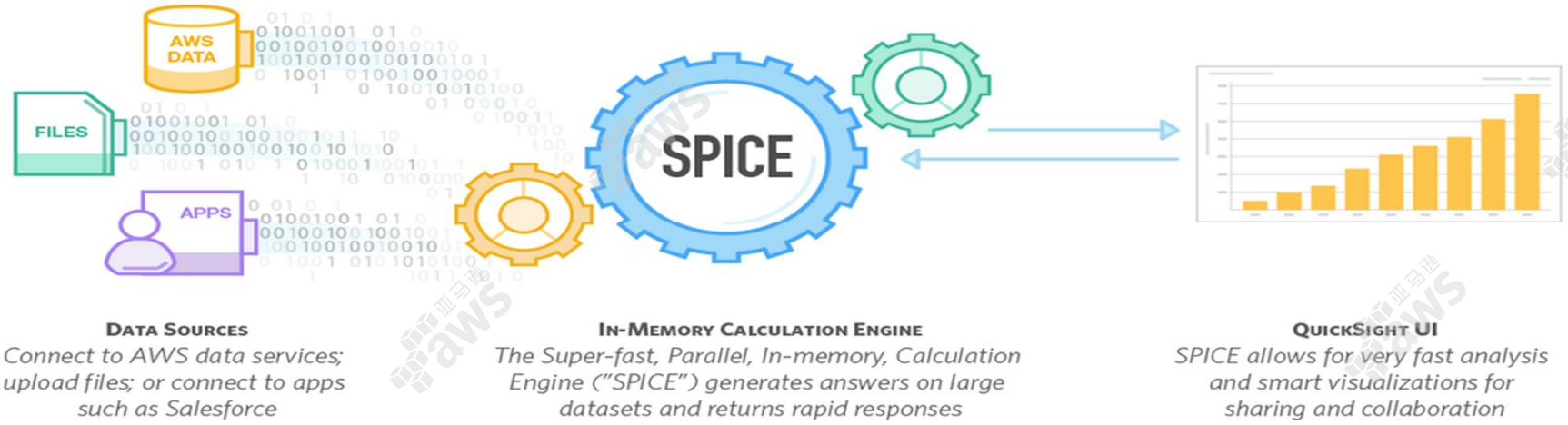

Amazon Redshift


第三方

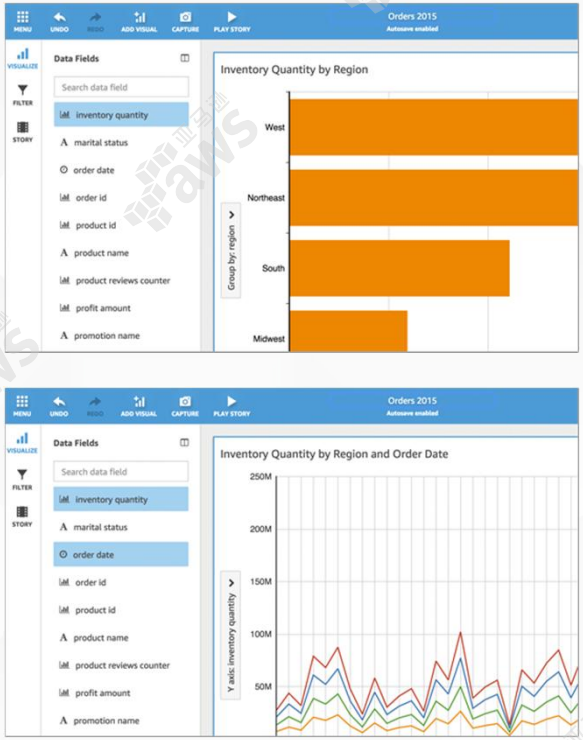
- 安全地发现和连接AWS数据服务
- 快速探查AWS数据源
  - 关系型数据库
  - NoSQL 数据库
  - Amazon EMR，Amazon S3和各类文件
  - 流数据源
- 简易地数据导入
- 自动探测数据类型

使用SPICE引擎快速获得业务见解

- **S**uper-fast, **P**arallel, **I**n-memory, **C**alculation **E**ngine
- 2到4倍列数据压缩
- 机器码生成编译过的查询
- 类SQL的语法
- 全托管服务－无需硬件和软件授权



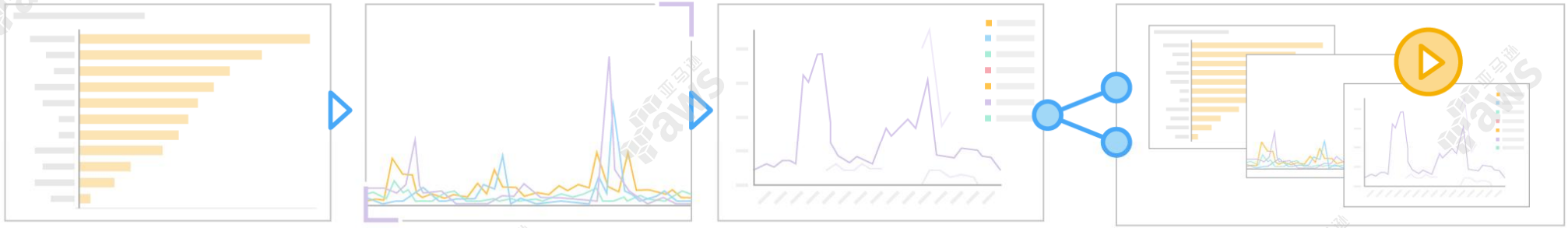
使用AutoGraph获得直观的视图



- 自动发现数据类型
- 生成优化查询
- 选择适合的展示图形类型
- 非常快速的响应

用数据讲故事

- 抓取关键的数据分析快照
- 创建一系列的分析
- 安全的共享
- 交互式数据探索
- 快速响应



Conduct your analysis and capture the visualizations as 'scenes' with annotations and comments

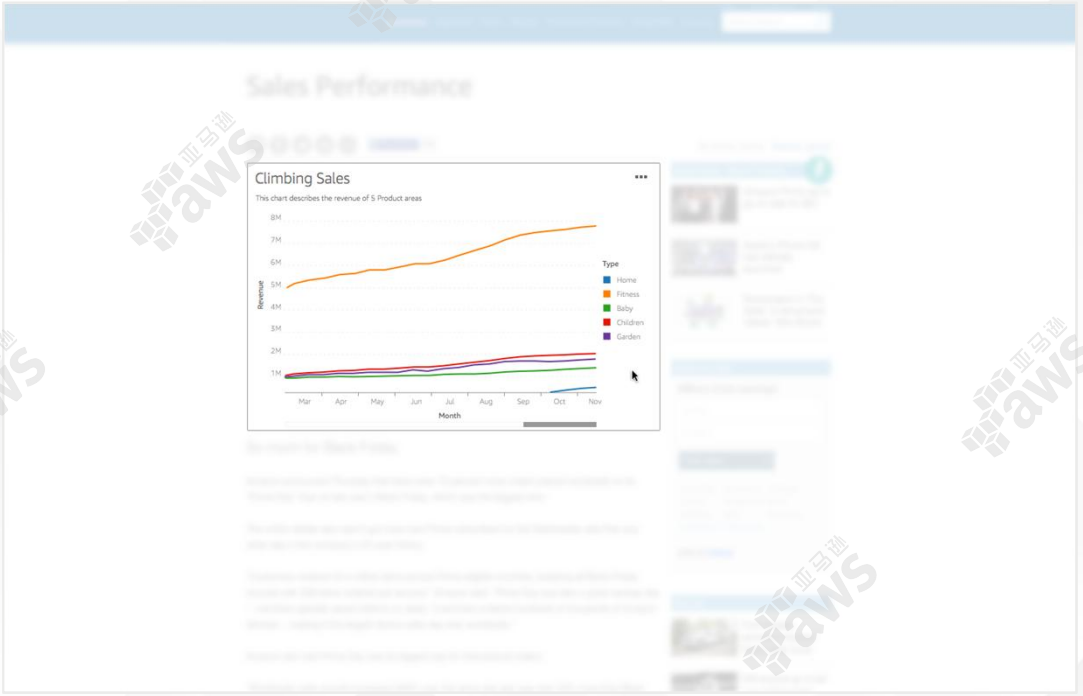
Share the scenes as a story to be played back anytime for further analysis

其他特性

移动设备支持

- iOS, Android
- 支持平板电脑

嵌入到现有应用



演示——分析销售数据

customer nar	customer typ	city	state	region	age r	gender	marital statu	product nam	category
Ⓐ	Ⓐ	Ⓐ	Ⓐ	Ⓐ	Ⓐ	Ⓐ	Ⓐ	Ⓐ	Ⓐ
Grant Emily	Individual	SUMMIT	NJ	Northeas	35 –	Male	Widowed	remark vases	Arts
Arnold Maria	Corporation	AVALON	TX	South	65+	Male	Divorced	suitable	Wellness
Russell Sara	Individual	CAIRO	NY	Northeas	55 –	Male	Divorced	globe toilet	Fitness
Rogers Fred	Individual	BUFFALO	NY	Northeas	18 –	Female	Divorced	privilege	Wellness
Lawson Laura	Corporation	BROWNS	IL	Midwest	18 –	Male	Widowed	swift plates	Home Decor
Bennett Fred	Individual	BARSTOW	CA	West	55 –	Male	Married	defend cake	Outdoor
Hudson Jimmy	Individual	GRANADA	CA	West	45 –	Female	Married	brilliant	Cycling
Mccoy Ernest	Individual	HILLSBORO	OR	West	35 –	Female	Single	loyal	Wireless
Murray Arthur	Corporation	MOLALLA	OR	West	18 –	Female	Divorced	actual plates	Outdoor/Sports
Carter Mark	Individual	OKANOGAN	WA	West	45 –	Male	Divorced	recall baking	Cycling
Rice Margaret	Corporation	STAFFORD	CT	Northeas	45 –	Male	Widowed	chill cake tins	Wellness
Freeman John	Corporation	GLENWOOD	CO	West	35 –	Male	Married	respect	Gardening
Taylor Nicole	Corporation	COLORADO	CO	West	25 –	Male	Single	swift soap	Gardening
James Heather	Corporation	CORVALLIS	OR	West	25 –	Male	Married	Wander	Gardening

Thank You