云端的数据湖

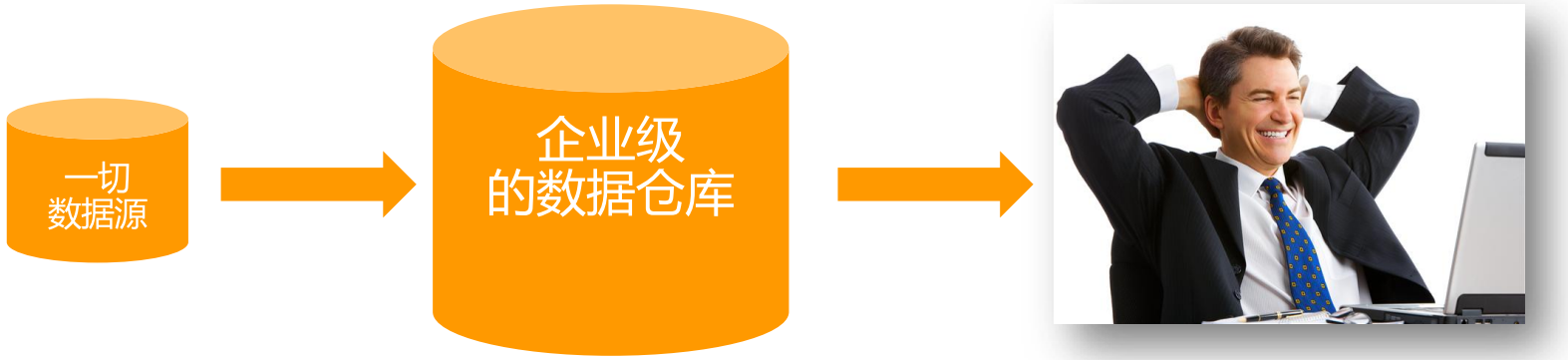
现代化的数据架构

张孝峰，AWS解决方案架构师

2018-1-25

aws

梦中的数据架构

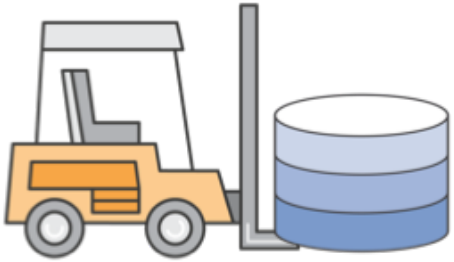


数据湖的优势 – 所有数据在一个地方

“我的数据储存在多个不同的地方，
那一份数据才是真实可信的呢？”

在一个集中的位置，
储存并分析来自所有来源的数据

数据湖的优势 – 快速提取

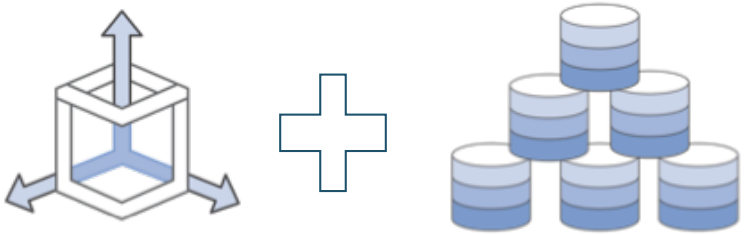


“如何快速从各种来源收集数据
并有效存储？”

快速提取数据，
而无需将其强制转换到范式中。

数据湖的优势 – 储存与计算分离

“如何扩展容量，
以应付持续增长的数据？”



将存储和计算分开，
可以根据需要缩放每个组件。

数据湖的优势 – 读取时范式化

“有没有办法将多个分析和处理框架应用于相同的数据？”

数据湖可以通过在读取时范式化来进行即时分析，而不是在写入时。

扩大使用者的范围

数据科学家

数据分析师

业务人员

第三方平台

自动化事件

1.更多的角色需要通过适当的工具访问数据

2.更多的系统需要链接到数据进行决策和过程自动化

3.用户需要能够查找信息并安全地访问它

业务数据呈指数级增长

1. 数据来自不同的来源，他们有不同的速度和规模

2. 需要把数据放在一起，打破传统的数据孤岛

3. 产生价值需要超过收集和分析的成本

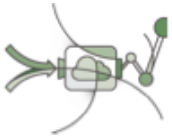
现代数据架构能够产生的价值

价值1 : 现代化的数据结构

- 洞察增强业务应用并创建新的数据服务

价值2 : 新的业务增长点

- 个性化，需要预测，风险评估



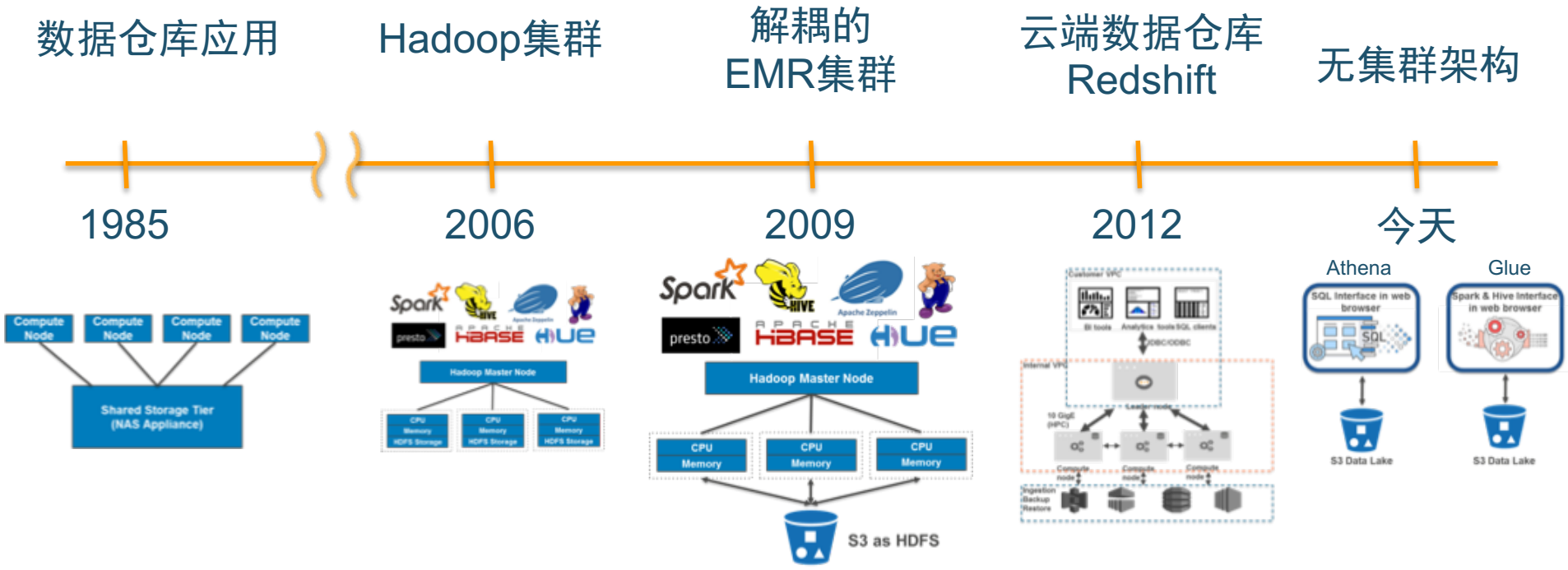
价值3 : 实时参与

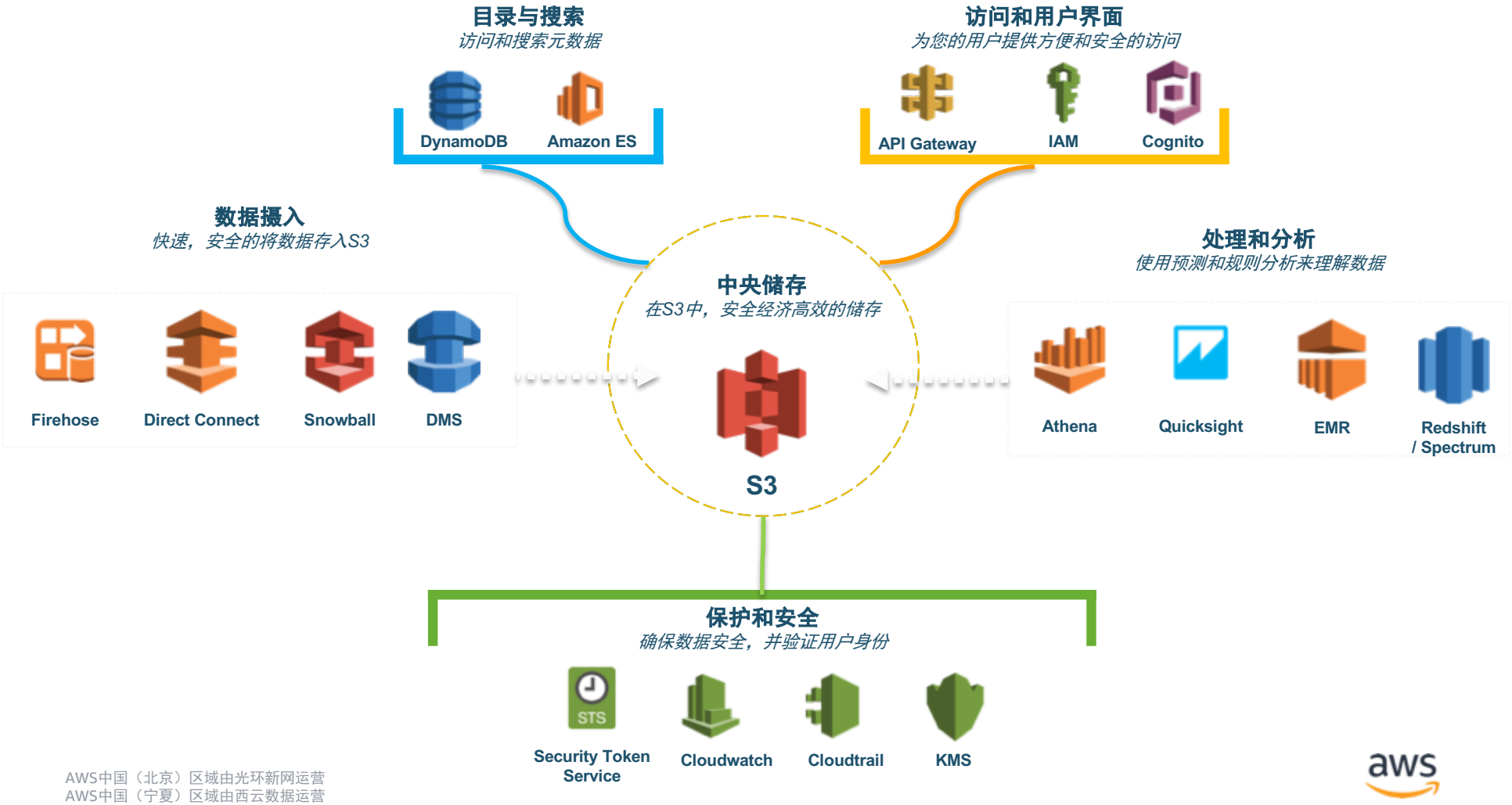
- 互动的客户体验，事件驱动的自动化，欺诈检测

价值4 : 自动扩展

- 业务流程和物理基础设施的自动化

数据分析平台技术的演变





AWS中国（北京）区域由光环新网运营
AWS中国（宁夏）区域由西云数据运营

为什么使用S3做为数据湖

持久性

提供高达11个9的数据持久性

可用性

提供99.99%可用性

高性能

- 多点上传
- 范围获取

易于使用

- REST API
- AWS SDKs
- Read-after-create持久性
- 事件通知、生命周期管理

扩展性

- 需要多少存多少
- 扩展储存和计算分离
- 没有最小使用量限制

集成性

- Amazon Redshift / Spectrum
- Amazon EMR
- Amazon Athena
- Amazon DynamoDB

入门：使用 Amazon EMR 分析大数据

http://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html

分析CloudFront的样例日志

s3://us-east-1.elasticmapreduce.samples/cloudfront/data/

1000KB sample data

```
2014-07-05  20:00:00    LHR3    4260    10.0.0.15   GET eabcd12345678.cloudfront.net    /test-image-1.jpeg  200 -   Mozilla/5.0%20(MacOS;%20U;%20Windows%20NT%205.1;
2014-07-05  20:00:00    MIA3    10  10.0.0.15   GET eabcd12345678.cloudfront.net    /test-image-1.jpeg  304 -   Mozilla/5.0%20(Linux;%20U;%20Windows%20NT%205.1;%20e
2014-07-05  20:00:00    MIA3    4252    10.0.0.15   GET eabcd12345678.cloudfront.net    /test-image-3.jpeg  200 -   Mozilla/5.0%20(Android;%20U;%20Windows%20NT%205.
2014-07-05  20:00:00    FRA2    4257    10.0.0.8    GET eabcd12345678.cloudfront.net    /test-image-2.jpeg  200 -   Mozilla/5.0%20(OSX;%20U;%20Windows%20NT%205.1;%2
2014-07-05  20:00:03    HKG1    4261    10.0.0.15   GET eabcd12345678.cloudfront.net    /test-image-2.jpeg  200 -   Mozilla/5.0%20(Windows;%20U;%20Windows%20NT%205.
2014-07-05  20:00:03    HKG1    4252    10.0.0.15   GET eabcd12345678.cloudfront.net    /test-image-1.jpeg  200 -   Mozilla/5.0%20(Windows;%20U;%20Windows%20NT%205.
2014-07-05  20:00:04    MIA3    4257    10.0.0.12   GET eabcd12345678.cloudfront.net    /test-image-3.jpeg  200 -   Mozilla/5.0%20(OSX;%20U;%20Windows%20NT%205.1;%2
2014-07-05  20:00:04    LAX1    4261    10.0.0.15   GET eabcd12345678.cloudfront.net    /test-image-1.jpeg  200 -   Mozilla/5.0%20(iOS;%20U;%20Windows%20NT%205.1;%2
```

更多模拟样例数据

我自己写了一个程序

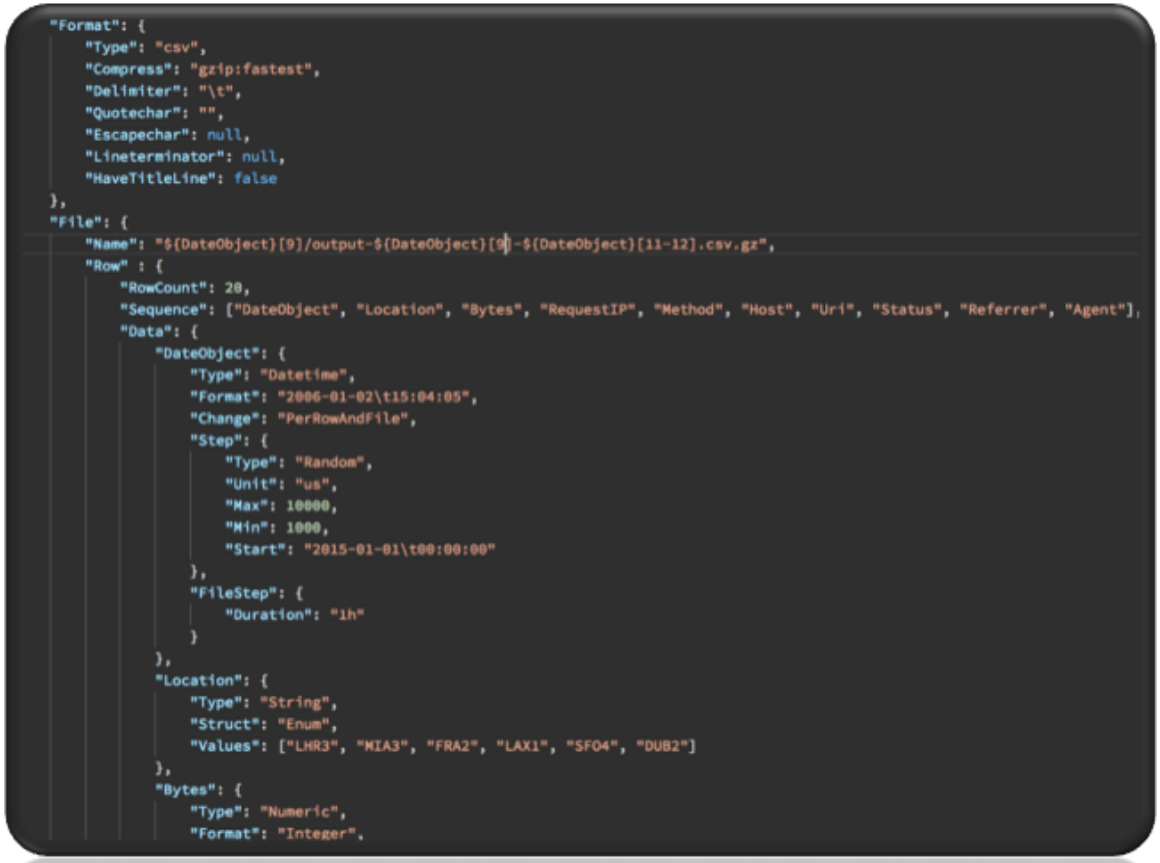
- https://github.com/tomcatzh/data-generator

使用Go语言

- Go去程实现高并发

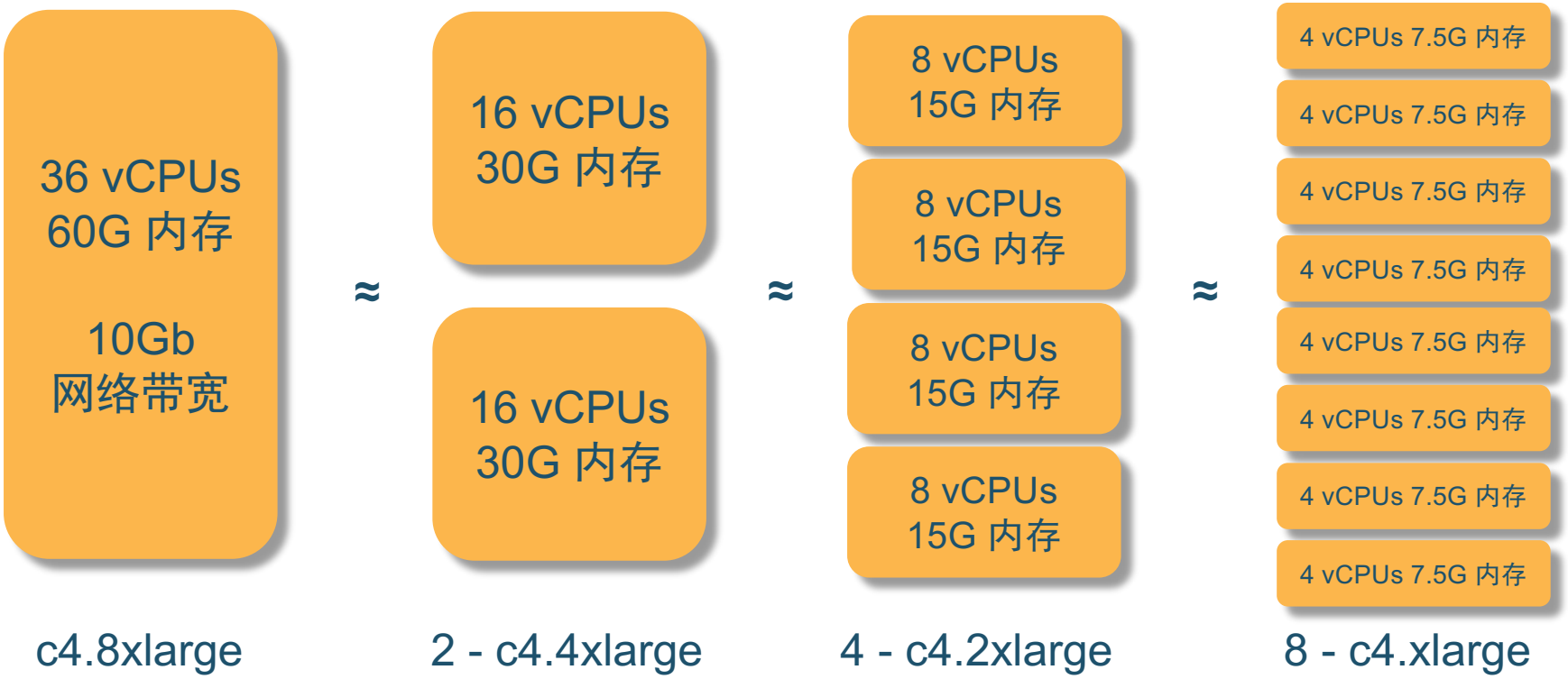
数据通过模板定制

充分随机



直接写入S3！

EC2实例大小



https://aws.amazon.com/ec2/instance-types/

那我能不能以超过5Gb的速度生成数据?

当然可以！横向扩展！

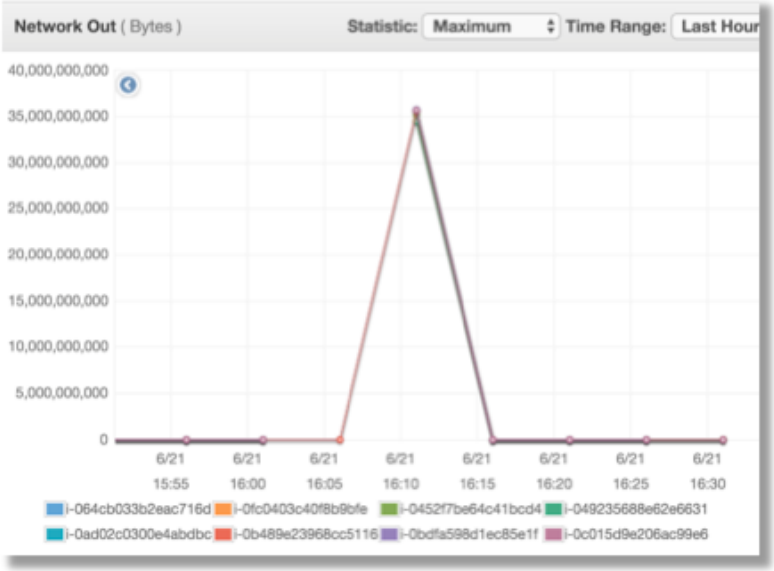
测试

- 8 x c4.4xlarge直写S3
- 期望能达到40 Gb/s

测试结果 - 8 x c4.4xlarge

8 x c4.4xlarge符合预期，接近40Gb/s（35.6Gb/s）

- 每个实例
  - 完成时间在3分59秒到4分10秒,
  - 数据生成速度在496到519MB/s
- 整个集群
  - 完成时间4m10s,
  - 整体数据生成速度3.88 GB/s

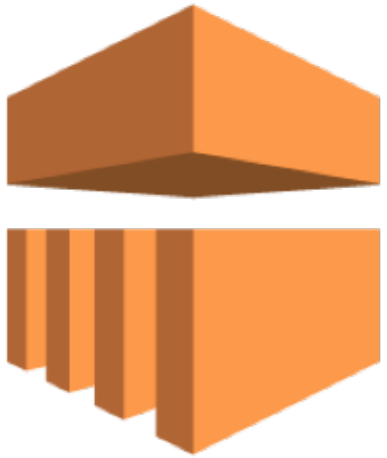


开始使用数据

读数据可以和写数据一样快吗？

硬件条件

17台c4.8xlarge核心节点，每台1000GB st1硬盘



硬件上限

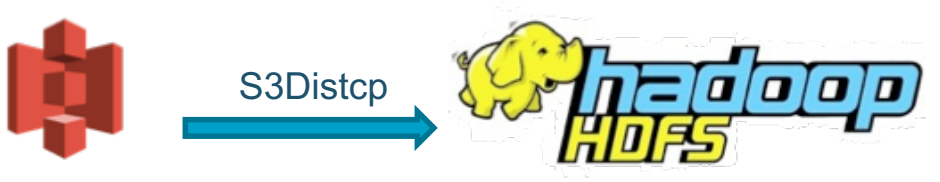
- CPU总数量612 vCPU
- 内存1TB
- 硬盘总容量 16.5TB
- 理论硬盘吞吐 680 / 4165MB (hdfs重复因子3)
- 理论网络带宽 85Gb/s

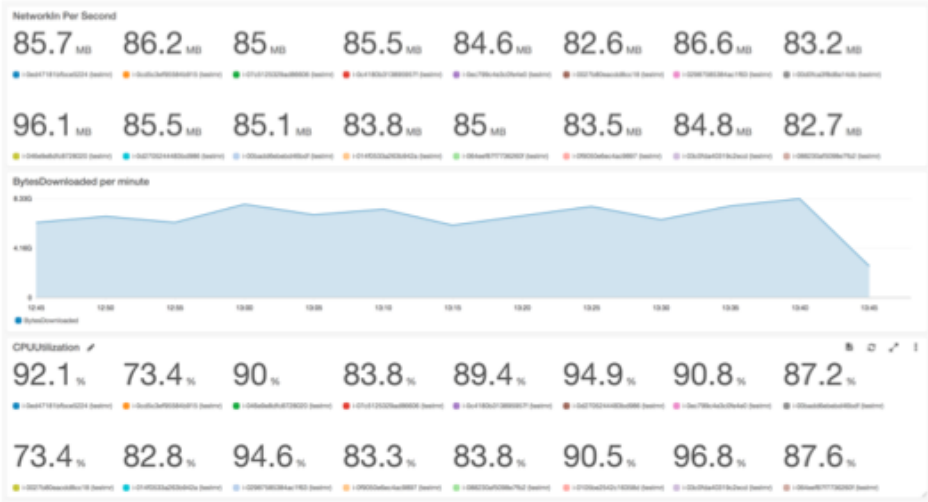
先来个小实验 - S3DistCp

从S3复制数据至HDFS

- 14分10秒完成
- S3下载速度10.48Gb/s

- HDFS的磁盘性能应满了
  - dfs.replication = 3
  - 实际写入带宽速率达3.93GB/s





http://docs.aws.amazon.com/zh_cn/emr/latest/ReleaseGuide/UsingEMR_s3distcp.html

重新做一下小实验 - S3DistCp

将dfs.replication设为1

- 4分36秒完成拷贝，32Gb/s

aws

测试语句 - 模拟两年，50亿行数据，1TB csv

```
SELECT os, COUNT(*) FROM cloudfront_log GROUP BY os
```

	dateobject	time	location	bytes	requestip	method	host	uri	status	referrer	os	browser	browserversion
1	2016-03-18	02:00:00	FRA2	830	10.107.235.179	GET	eabcd12345678.cloudfront.net	/test-image-3.jpeg	401	-	Linux	Chrome	3.0.9
2	2016-03-18	02:00:00	LHR3	3125	10.12.92.248	POST	eabcd12345678.cloudfront.net	/test-image-2.jpeg	304	-	iOS	Lynx	3.0.9
3	2016-03-18	02:00:00	DUB2	7592	10.237.194.174	GET	eabcd12345678.cloudfront.net	/test-image-1.jpeg	401	-	Linux	Chrome	3.0.9
4	2016-03-18	02:00:00	LAX1	7080	10.185.193.6	GET	eabcd12345678.cloudfront.net	/test-image-3.jpeg	404	-	Windows	Firefox	3.0.9
5	2016-03-18	02:00:00	LHR3	8082	10.245.114.213	GET	eabcd12345678.cloudfront.net	/test-image-1.jpeg	500	-	MacOS	IE	3.0.9
6	2016-03-18	02:00:00	LHR3	8269	10.133.174.190	POST	eabcd12345678.cloudfront.net	/test-image-3.jpeg	200	-	Windows	Opera	3.0.9
7	2016-03-18	02:00:00	FRA2	9400	10.154.49.232	POST	eabcd12345678.cloudfront.net	/test-image-2.jpeg	404	-	OSX	Chrome	3.0.9
8	2016-03-18	02:00:00	LAX1	9274	10.159.40.106	GET	eabcd12345678.cloudfront.net	/test-image-3.jpeg	304	-	iOS	Lynx	3.0.9
9	2016-03-18	02:00:00	MIA3	2724	10.157.57.120	GET	eabcd12345678.cloudfront.net	/test-image-3.jpeg	401	-	Windows	Opera	3.0.9
10	2016-03-18	02:00:00	DUB2	1797	10.90.134.198	GET	eabcd12345678.cloudfront.net	/test-image-1.jpeg	304	-	Linux	Chrome	3.0.9

Hive查询两天数据

总体数据量3124.66MB

S3

- 12分14秒完成查询

HDFS

- 9分54秒完成查询

Hive查询两年数据

总体数据量为1113GB

S3

- 16小时43分54秒完成查询
- 扫描速度18.98MB/s

HDFS

- 17小时15分41秒完成查询
- 扫描速度18.35MB/s

使用压缩数据

Hive支持数据压缩

重新生成数据

- 同样的分区，同样的两年条目数
- 使用gzip2的最快压缩等级
- 104.2GB，压缩率91%

EMR Hive查询

- 16小时41分29秒完成查询

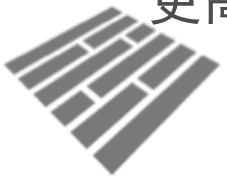
更适合大数据的格式 - Parquet

需要进行主动转换

- http://docs.aws.amazon.com/athena/latest/ug/convert-to-columnar.html

两年数据转换时间

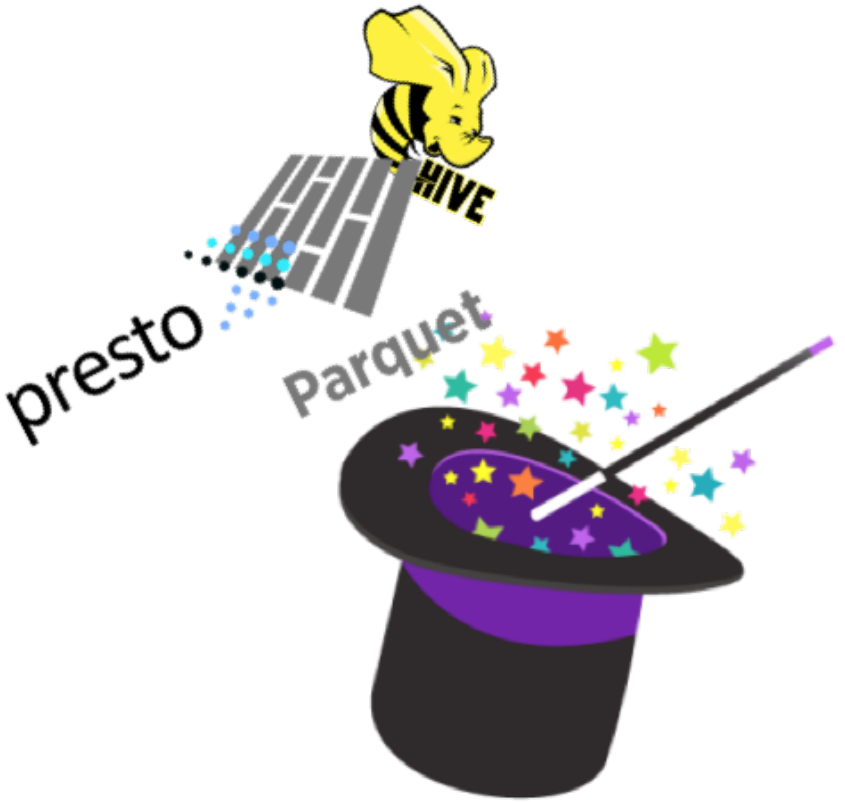
- 耗费15小时5分50秒
- 转换后大小109.8GB，压缩率90%（向量化，使用snappy可以有更高的压缩率）



更适合查询的数据格式 - Parquet

EMR Hive查询

- 用时88秒



使用Athena进行查询

Athena查询

- 用时8.75秒，扫描量 2.02GB
- 成本 $0.01

aws

分区

Hive世界的分区很暴力，但很有效

分区效果明显

```
SELECT item_date,
        entry_id,
        entry_t,
        region,
        cast(sum(entry_sv) AS DECIMAL(19,
        6)) AS cnt_entry_sv
FROM "item"
LEFT JOIN "user"
    ON m_id=userid
WHERE entry_id in(8,7,5,1,9,6,14,10)
       AND year='2017'
       AND month='10'
GROUP BY  item_date ,entry_t,entry_id,region
ORDER BY  item_date DESC limit 100
```

- item表 一千亿条 csv gzip 2T
- user表 五千万条 csv gzip 1.5G

使用Glue自动建立分区

查询时常 8.47秒
数据扫描量 10.19GB
成本 $0.05

S3可以作为大数据的热储存

高吞吐（优于HDFS）

安全可控

免维护

存放在HDFS

1PB数据

存放在S3

复制因子 = 3
3 PB

磁盘预留空间 25%
4 PB

1
PB

单价
ST1: $0.045 / 每月GB

总价
$ 188,743.68 / 月

单价
S3: 约 $0.02155 / 每月GB

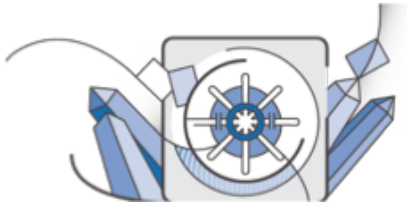
总价
$ 22,067.2 / 月

价格以美国东部（弗吉尼亚北部）为例 https://aws.amazon.com/s3/pricing/

你可以省更多

标准存储

标准低频率访问存储



Glacier 存储

更便宜！

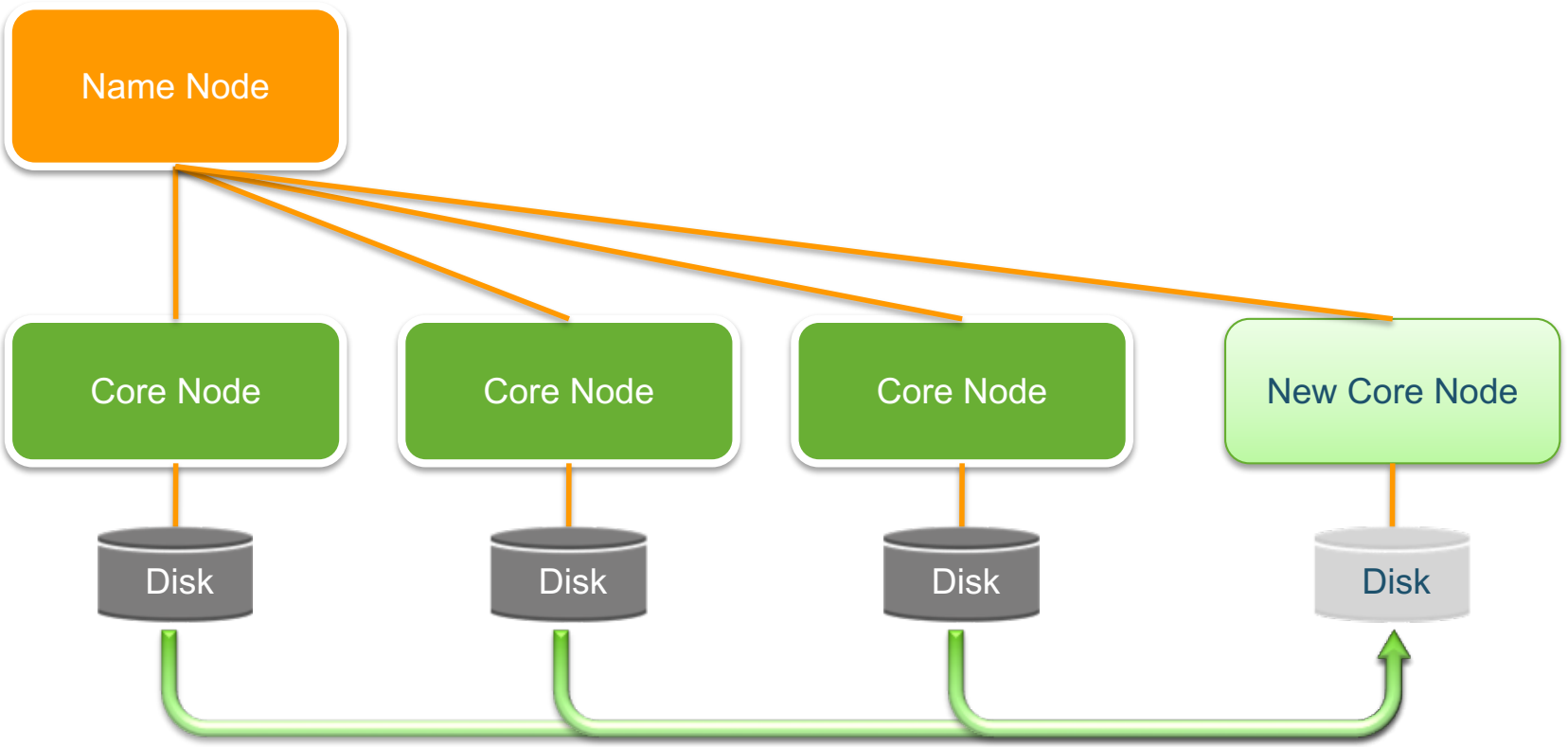
计算能力与储存解耦

计算能力

储存能力

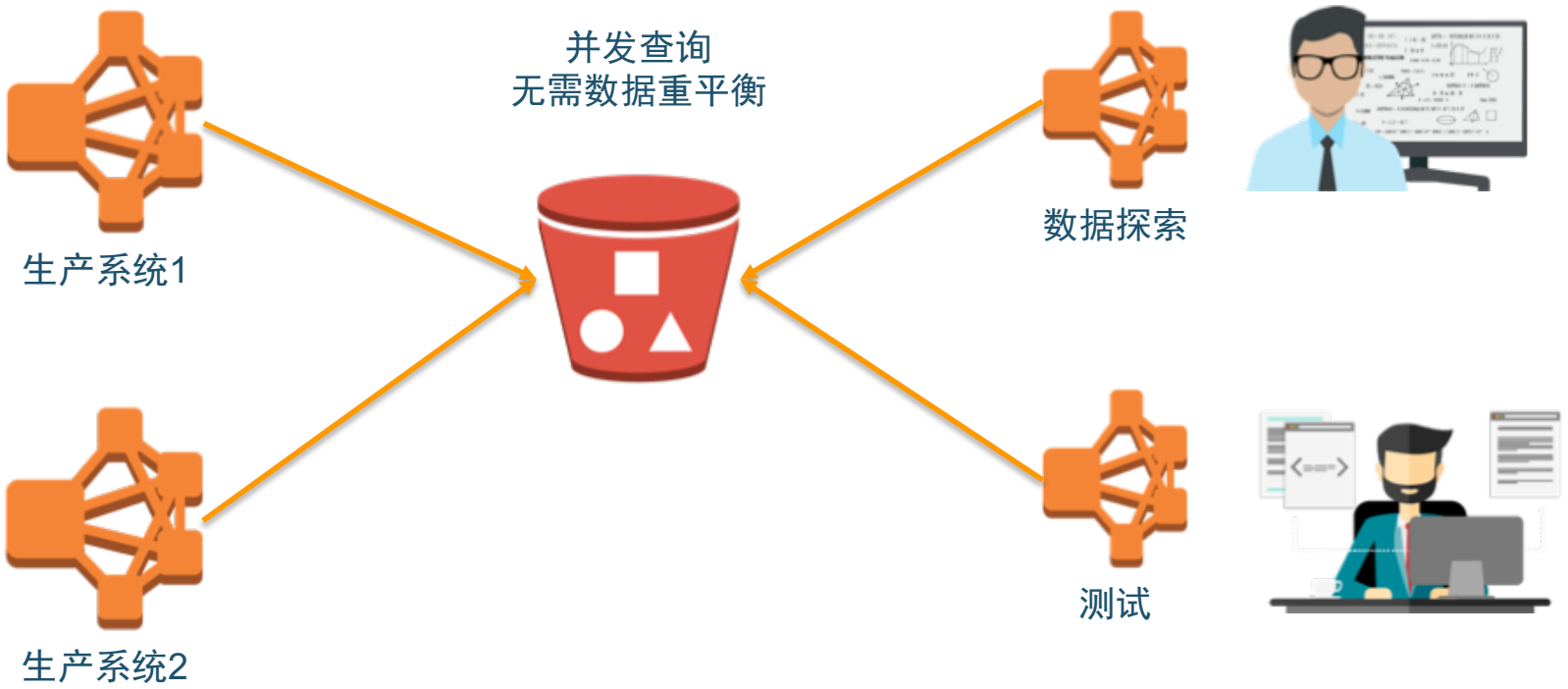
aws

Hadoop HDFS数据的重平衡



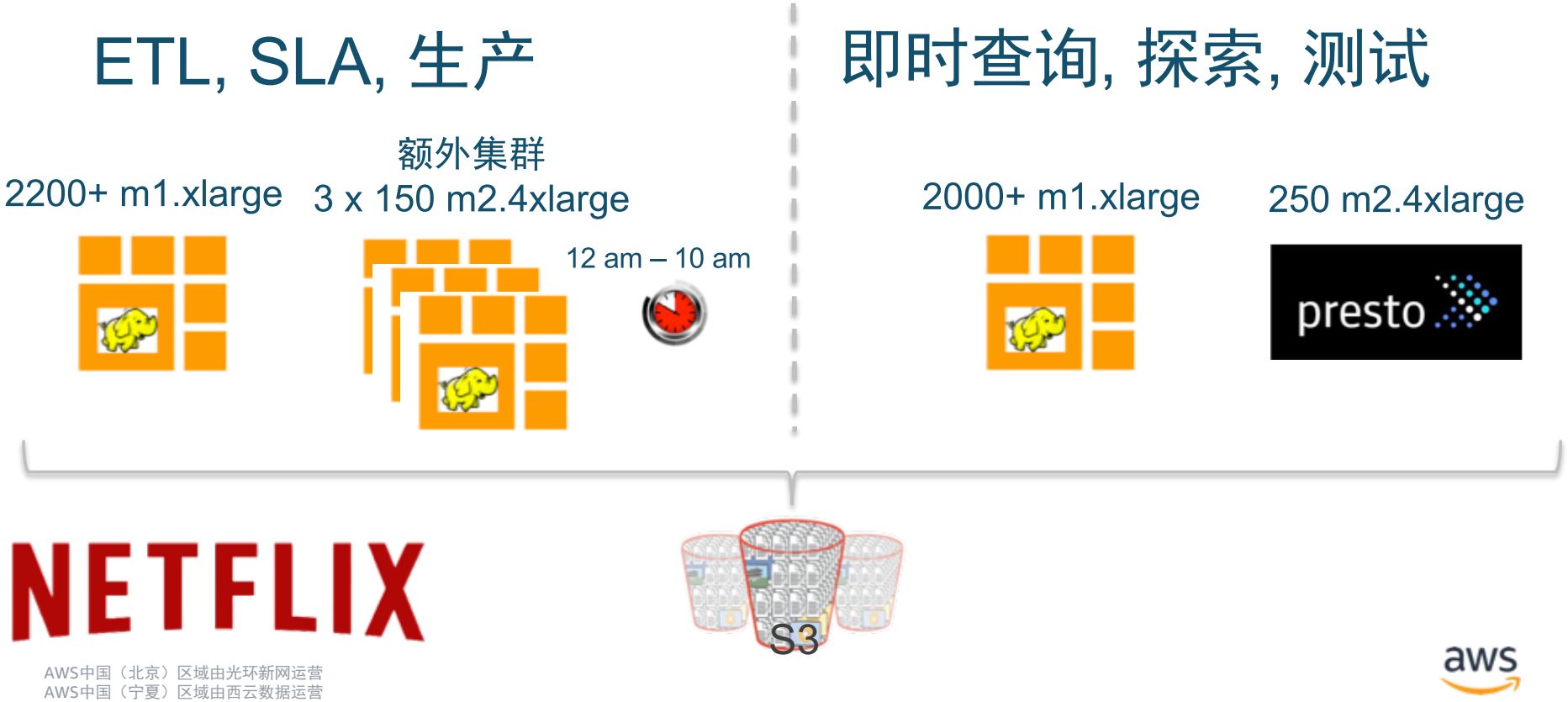
数据重平衡

S3支持多个EMR集群同时查询同一批数据



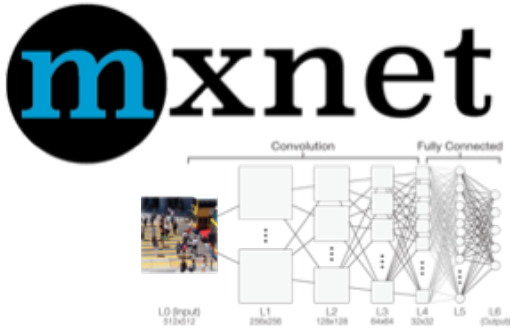

House of Cards
2013 TV-MA 1 Season HD 5.1
Sharks gliding ominously beneath the surface of the water? They're a lot less menacing than this Congressman.
This winner of three Emmys, including Outstanding Directing for David Fincher, stars Kevin Spacey and Robin Wright.
NETFLIX
NETFLIX
Because you watched Orange Is the New Black
Breaking Bad
A NETFLIX ORIGINAL SERIES
HOUSE of CARDS
the L word
THE LEAGUE
IT'S ALWAYS SUNNY IN PHILADELPHIA
New Girl
Because you watched Red Lights
NEWMAN SARANDON HACKMAN TWILIGHT
6 SOULS
TALL
THE ELECTRIC MIST
THE FACTORY
INTRUDERS

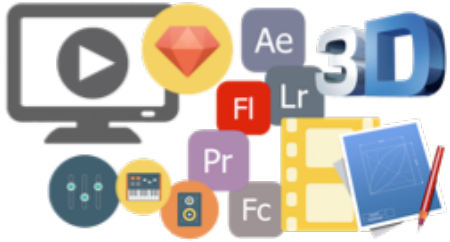
Netflix使用S3作为可扩展的数据架构

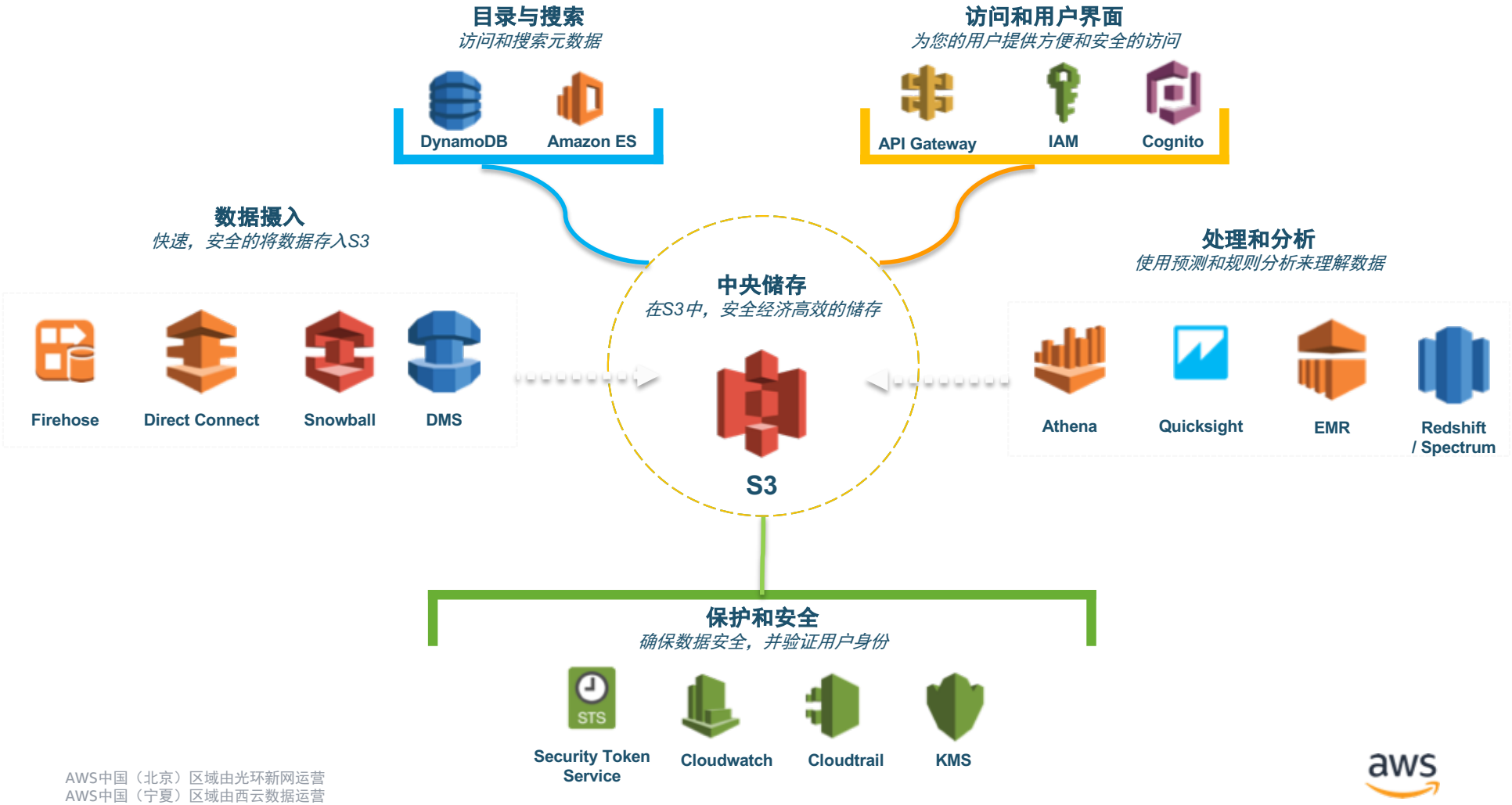


AWS中国（北京）区域由光环新网运营
AWS中国（宁夏）区域由西云数据运营

更多的大数据可以使用S3







AWS中国（北京）区域由光环新网运营
AWS中国（宁夏）区域由西云数据运营

Thank you!

现在就可以动起手来，构造你第一个数据湖