DAT201

Introduction to Amazon Redshift

Pavan Pothukuchi, Amazon Redshift
Nam Nguyen, RetailMeNot

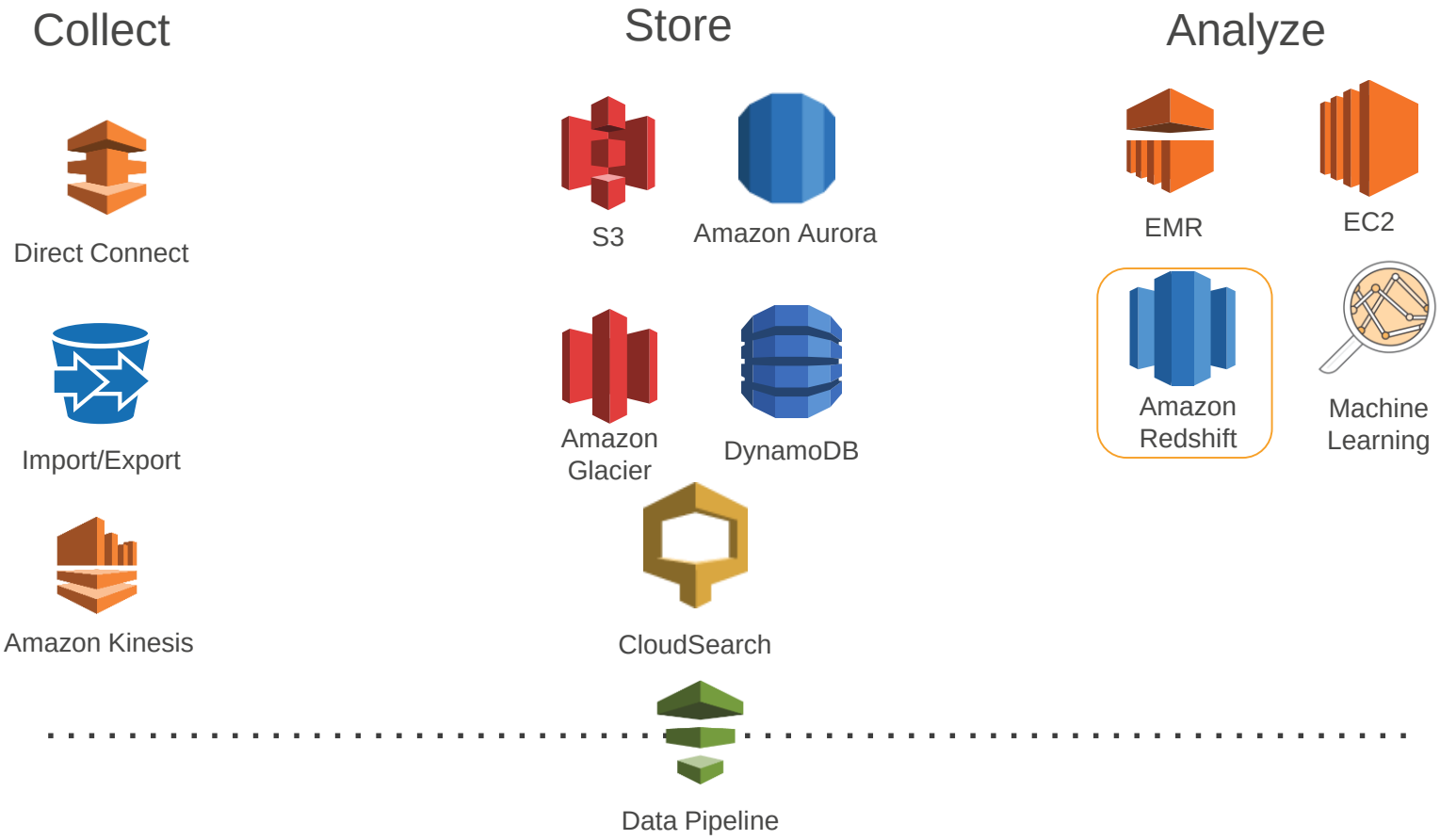
October 2015

amazon web services

What to expect from the session

- Amazon Redshift – What and Why
- Benefits
- Use cases
- Amazon Redshift at RetailMeNot
- Q&A

AWS big data portfolio



Amazon Redshift

a lot faster
a lot simpler
a lot cheaper

Relational data warehouse

Massively parallel; Petabyte scale

Fully managed

HDD and SSD Platforms

$1,000/TB/Year; starts at $0.25/hour

The legacy view of data warehousing ...

Global 2,000 companies

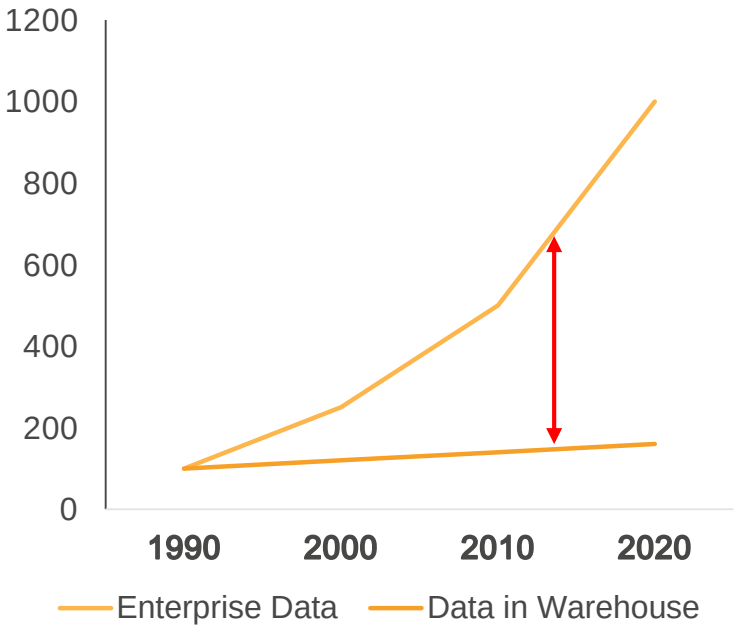
Sell to central IT

Multi-year commitment

Multi-year deployments

Multi-million dollar deals

… Leads to dark data



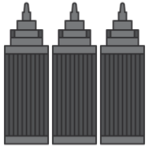
This is a narrow view

Small companies also have big data

(mobile, social, gaming, adtech, IoT)

Long cycles, high costs, administrative complexity all stifle innovation

The Amazon Redshift view of data warehousing



Enterprise

10x cheaper

Easy to provision

Higher DBA productivity

Big Data

10x faster

No programming

Easily leverage BI tools, Hadoop, Machine Learning, Streaming

SaaS

Analysis in-line with process flows

Pay as you go, grow as you need

Managed availability & DR

Selected Amazon Redshift customers

Amazon Redshift architecture

Leader Node

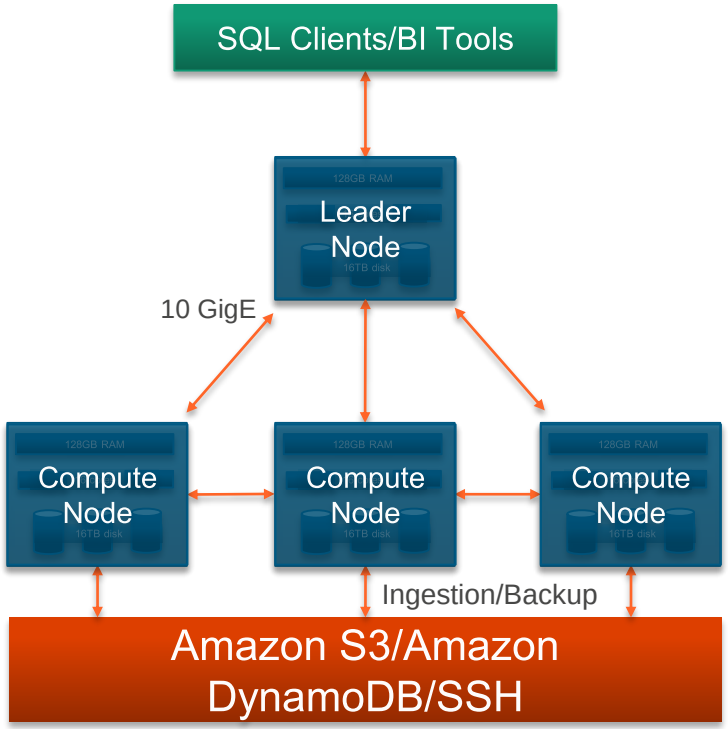
- Simple SQL end point
- Stores metadata
- Optimizes query plan
- Coordinates query execution

Compute Nodes

- Local columnar storage
- Parallel/distributed execution of all queries, loads, backups, restores, resizes

Start at just $0.25/hour, grow to 2 PB (compressed)

- DC1: SSD; scale from 160 GB to 326 TB
- DS2: HDD; scale from 2 TB to 2 PB



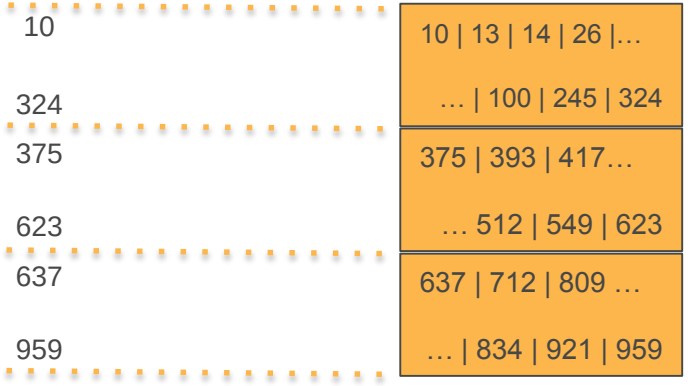
Benefit #1: Amazon Redshift is fast

Dramatically less I/O

- Column storage
- Data compression
- Zone maps
- Direct-attached storage
- Large data block sizes

```
analyze compression listing;

  Table   |     Column      | Encoding
----------+-----------------+----------
 listing | listid          | delta
 listing | sellerid        | delta32k
 listing | eventid         | delta32k
 listing | dateid          | bytedict
 listing | numtickets      | bytedict
 listing | priceperticket  | delta32k
 listing | totalprice      | mostly32
 listing | listtime        | raw
```

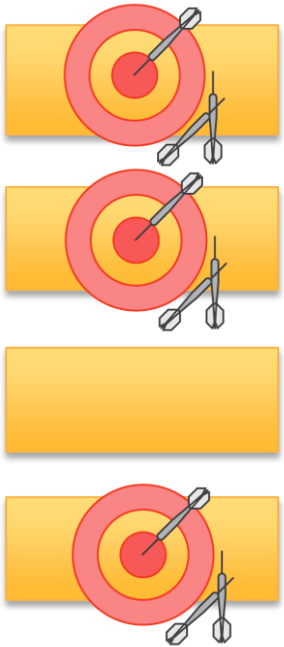


Benefit #1: Amazon Redshift is fast

Sort Keys and Zone Maps

SELECT COUNT(*) FROM LOGS WHERE DATE = '09-JUNE-2013'

Unsorted Table



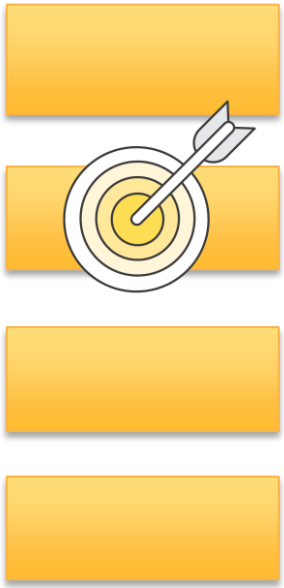
MIN: 01-JUNE-2013
MAX: 20-JUNE-2013

MIN: 08-JUNE-2013
MAX: 30-JUNE-2013

MIN: 12-JUNE-2013
MAX: 20-JUNE-2013

MIN: 02-JUNE-2013
MAX: 25-JUNE-2013

Sorted By Date



MIN: 01-JUNE-2013
MAX: 06-JUNE-2013

MIN: 07-JUNE-2013
MAX: 12-JUNE-2013

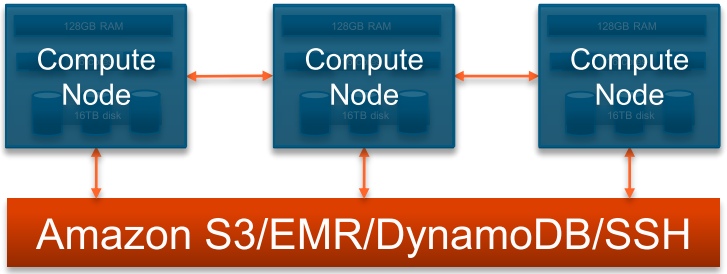
MIN: 13-JUNE-2013
MAX: 18-JUNE-2013

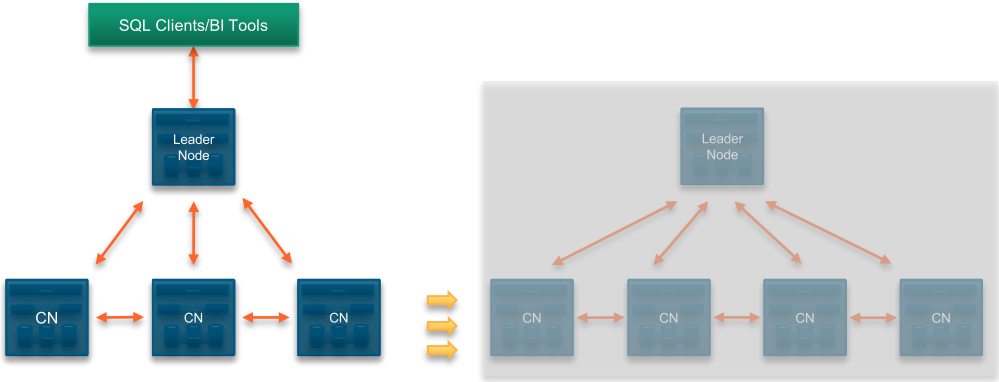
MIN: 19-JUNE-2013
MAX: 24-JUNE-2013

Benefit #1: Amazon Redshift is fast

Parallel and Distributed

- Query
- Load
- Export
- Backup
- Restore
- Resize

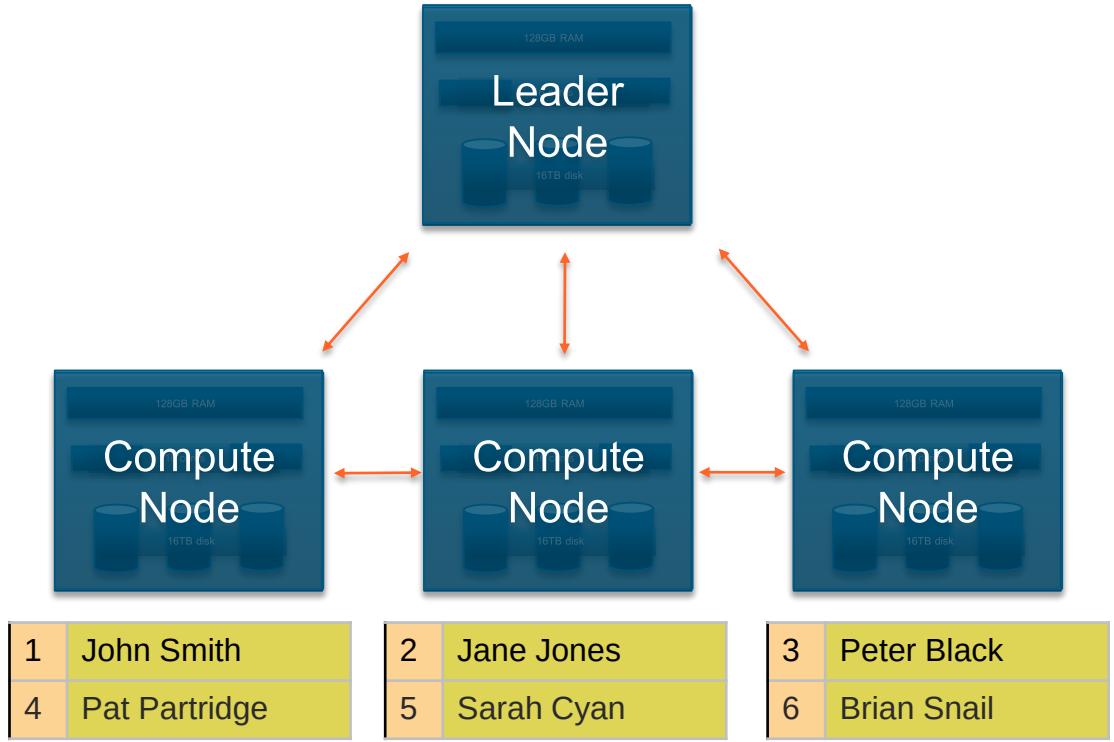




Benefit #1: Amazon Redshift is fast

Distribution Keys

ID	Name
1	John Smith
2	Jane Jones
3	Peter Black
4	Pat Partridge
5	Sarah Cyan
6	Brian Snail



Benefit #1: Amazon Redshift is fast

H/W optimized for I/O intensive workloads, 4GB/sec/node

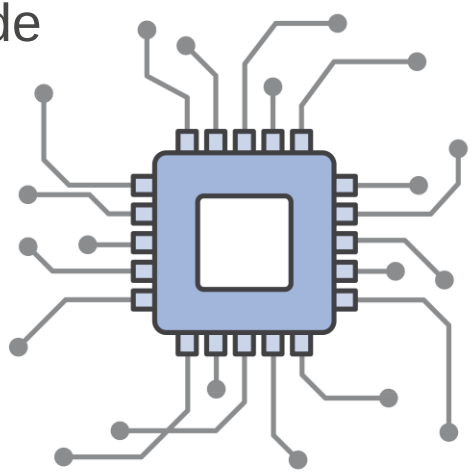
Enhanced networking, over 1M packets/sec/node

Choice of storage type, instance size

Regular cadence of auto-patched improvements

Example: Our new Dense Storage (HDD) instance type

- Improved memory 2x, compute 2x, disk throughput 1.5x
- Cost: same as our prior generation !



Benefit #2: Amazon Redshift is inexpensive

DS2 (HDD)	Price Per Hour for DW1.XL Single Node	Effective Annual Price per TB compressed
On-Demand	$ 0.850	$ 3,725
1 Year Reservation	$ 0.500	$ 2,190
3 Year Reservation	$ 0.228	$ 999

DC1 (SSD)	Price Per Hour for DW2.L Single Node	Effective Annual Price per TB compressed
On-Demand	$ 0.250	$ 13,690
1 Year Reservation	$ 0.161	$ 8,795
3 Year Reservation	$ 0.100	$ 5,500

Pricing is simple

Number of nodes x price/hour

No charge for leader node

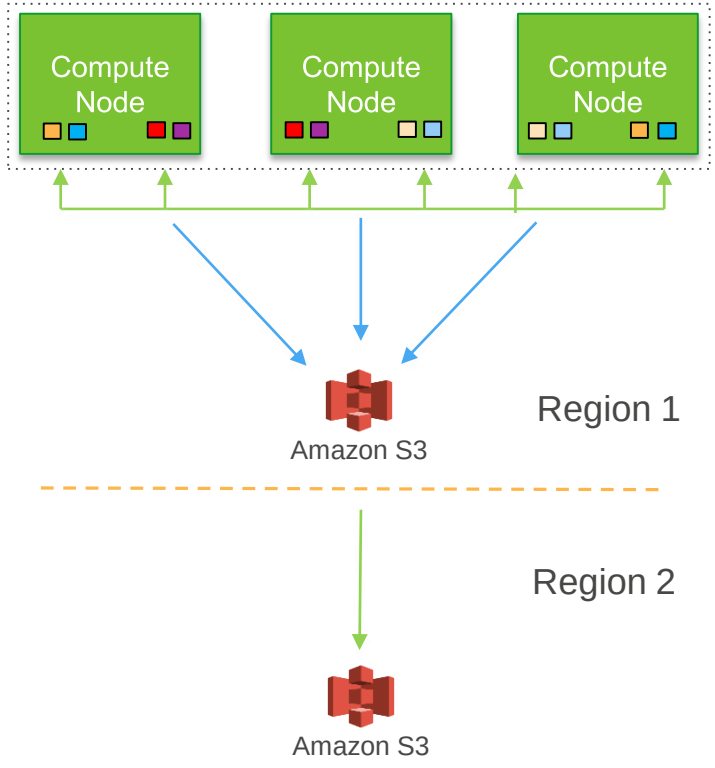
No up front costs

Pay as you go

Benefit #3: Amazon Redshift is fully managed

Continuous/incremental backups

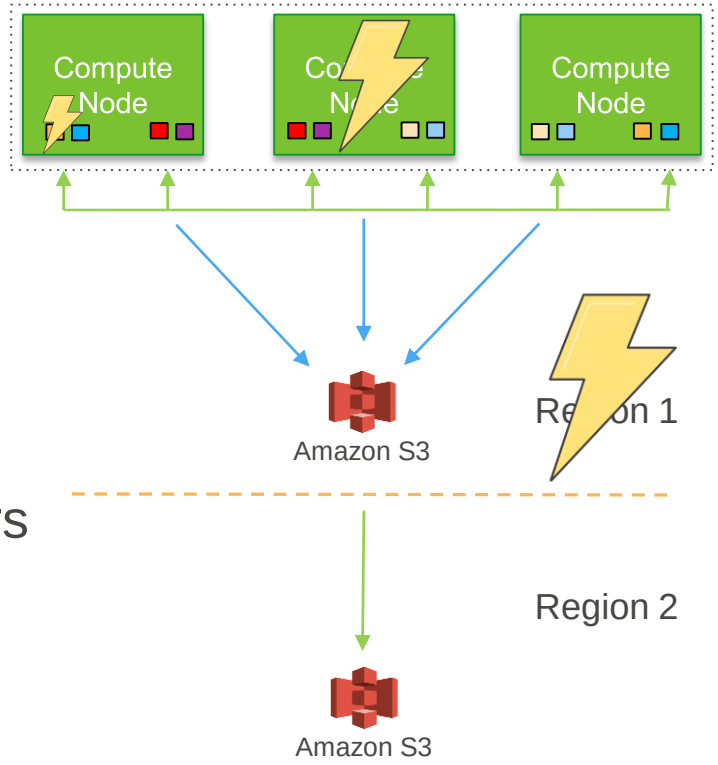
- Multiple copies within cluster
- Continuous and incremental backups to S3
- Continuous and incremental backups across regions
- Streaming restore



Benefit #3: Amazon Redshift is fully managed

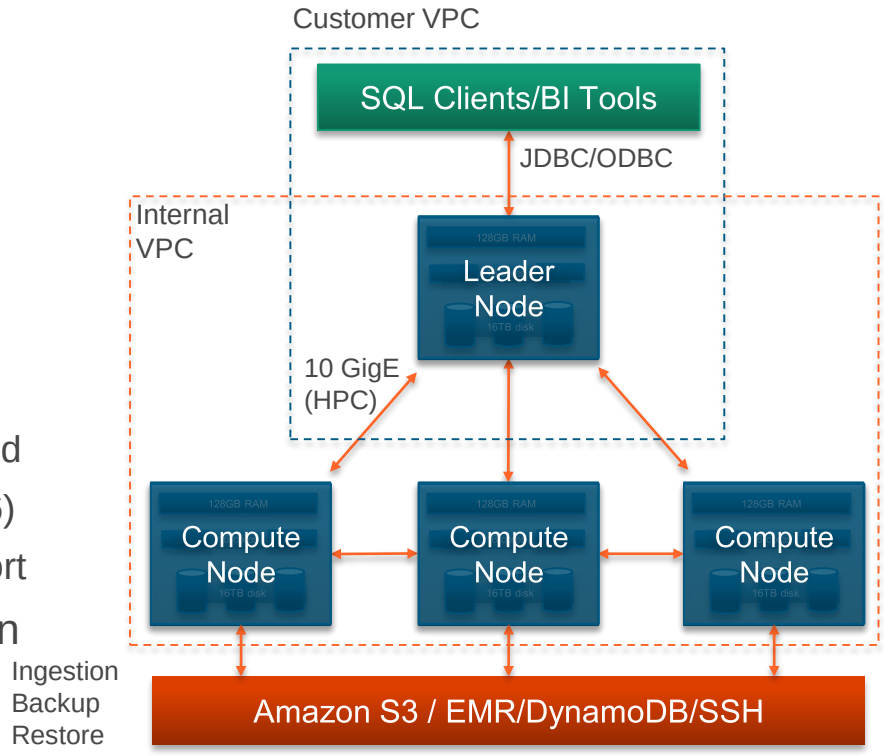
Fault tolerance

- Disk failures
- Node failures
- Network failures
- Availability Zone/Region level disasters



Benefit #4: Security is built-in

- Load encrypted from S3
- SSL to secure data in transit
  - ECDHE perfect forward security
- Amazon VPC for network isolation
- Encryption to secure data at rest
  - All blocks on disks & in Amazon S3 encrypted
  - Block key, Cluster key, Master key (AES-256)
  - On-premises HSM & AWS CloudHSM support
- Audit logging and AWS CloudTrail integration
- SOC 1/2/3, PCI-DSS, FedRAMP, BAA

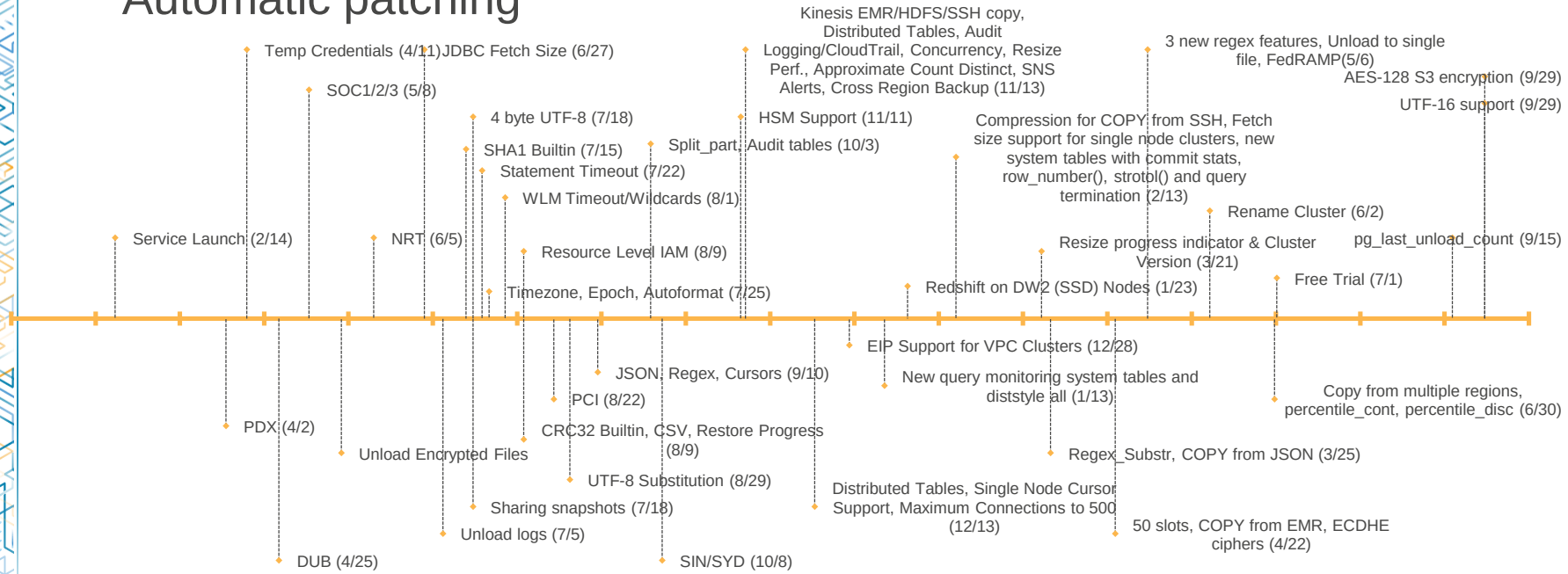


Benefit #5: We innovate quickly

Well over 100 new features added since launch

Release every two weeks

Automatic patching



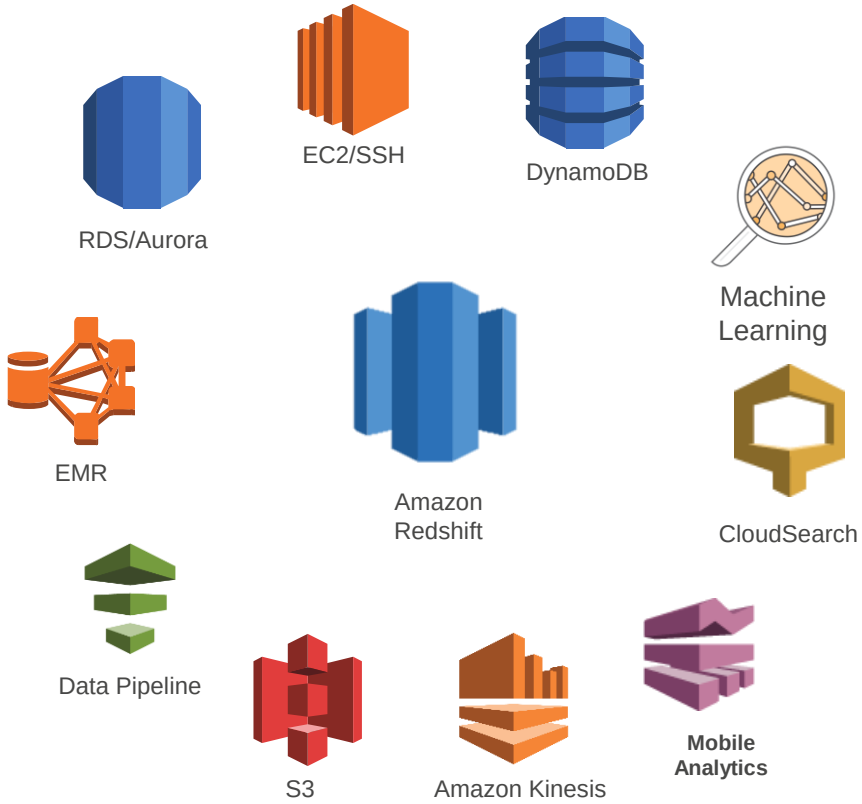
Benefit #6: Amazon Redshift is powerful

- Approximate functions
- User defined functions
- Machine Learning
- Data Science

Amazon ML

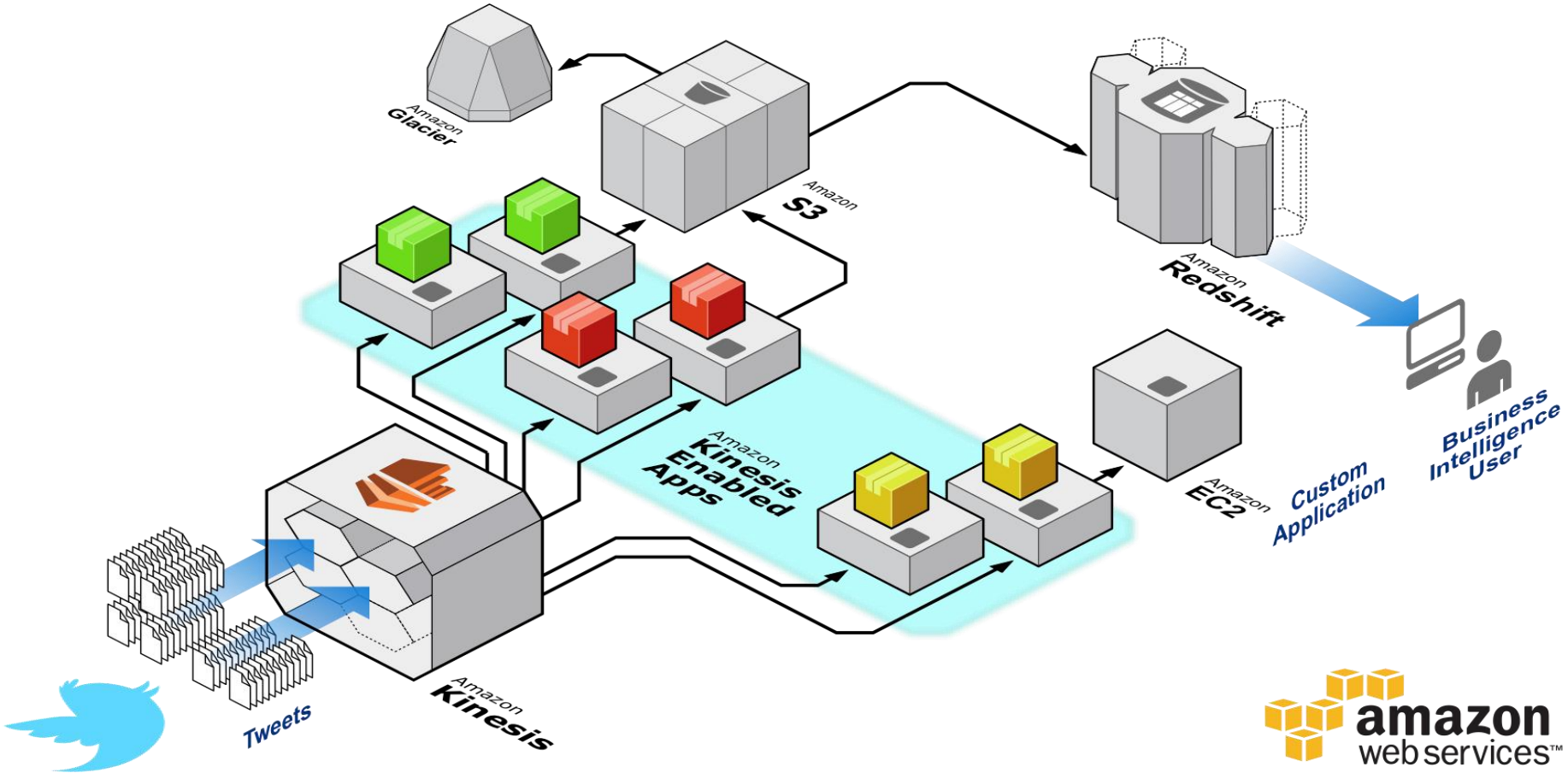
Benefit #7: Amazon Redshift has a large ecosystem

Benefit #8: Service oriented architecture

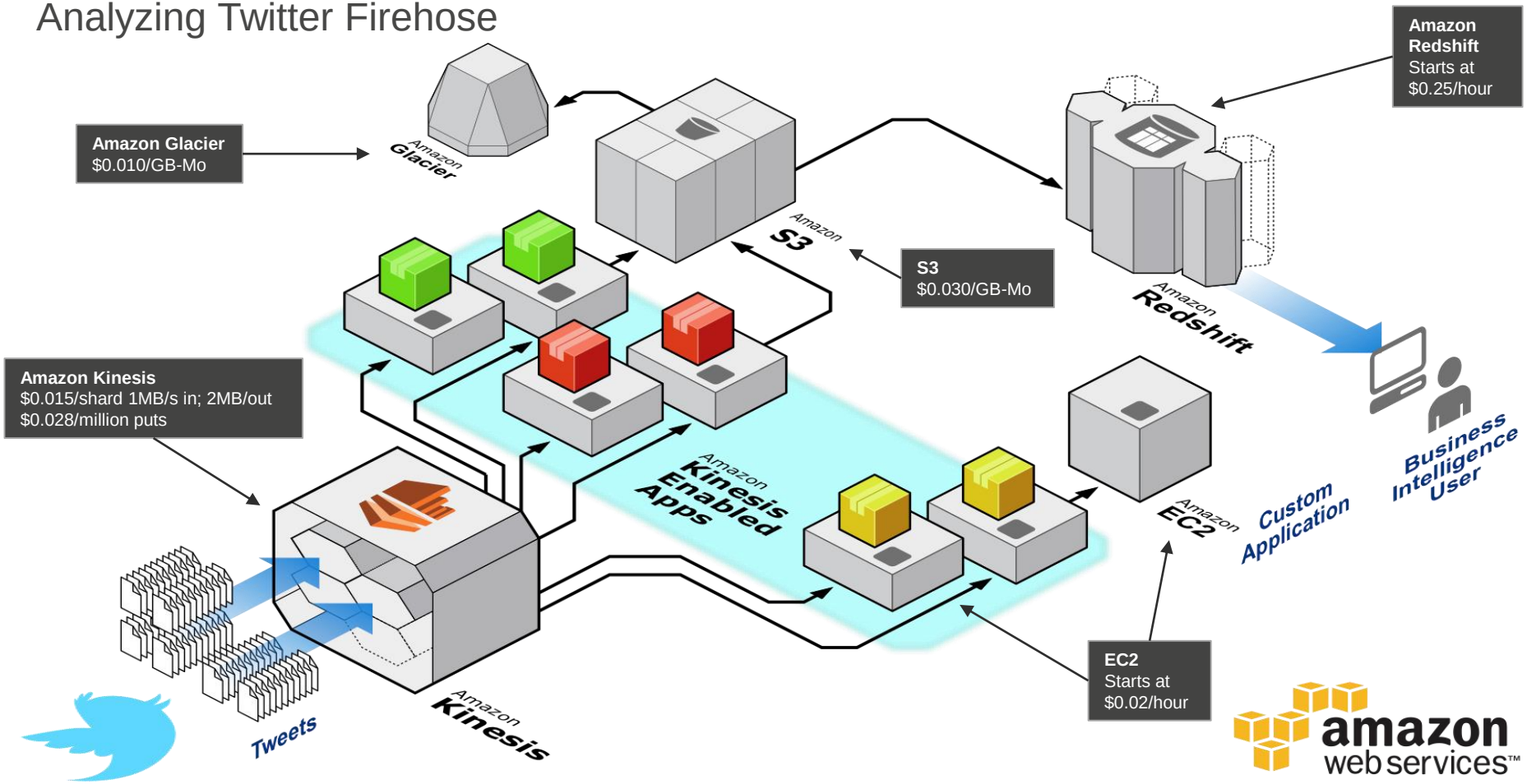


Use cases

Analyzing Twitter Firehose



Analyzing Twitter Firehose



Data warehouses can be inexpensive and powerful

500MM tweets/day = ~ 5,800 tweets/sec

2k/tweet is ~12MB/sec (~1TB/day)

$0.015/hour per shard, $0.028/million PUTS

Amazon Kinesis cost is $0.765/hour

Amazon Redshift cost is $0.850/hour (for a 2TB node)

S3 cost is $1.28/hour (no compression)

Total: $2.895/hour

Data warehouses can be inexpensive and powerful

Use only the services you need

Scale only the services you need

Pay for what you use

~40% discount with 1 year commitment

~70% discounts with 3 year commitment

Amazon.com – Weblog analysis

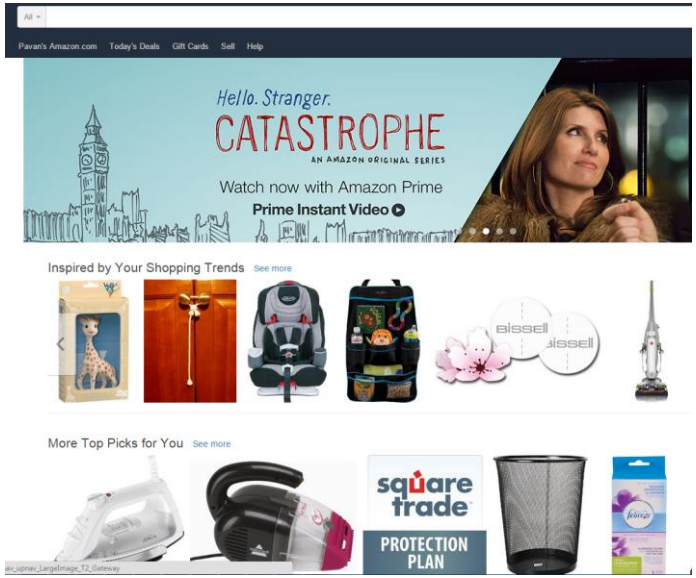
Web log analysis for Amazon.com

- 1PB+ workload, 2TB/day, growing 67% YoY
- Largest table: 400 TB

Want to understand customer behavior

Solution

- Legacy DW—query across 1 week/hr.
- Hadoop—query across 1 month/hr.



Data warehouses can be

fast

and

simple

Query 15 months of data (1PB) in 14 minutes

Load 5B rows in 10 minutes

21B rows joined with 10B rows – 3 days (Hive) to 2 hours

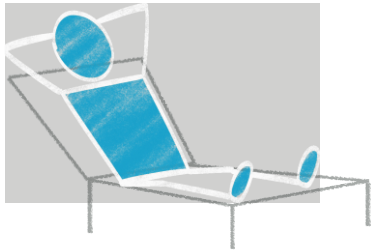
Load pipeline: 90 hours (Oracle) to 8 hours

64 clusters

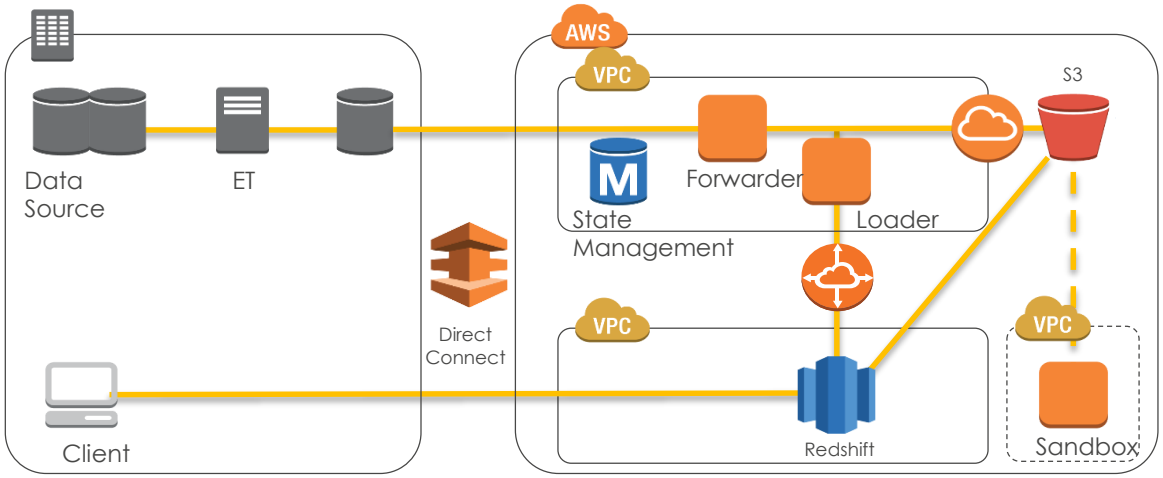
800 total nodes

13PB provisioned storage

2 DBAs



NTT Docomo – Mobile usage analysis



Petabytes of data generated by many cell phone towers

Hard to scale, expensive

Needed a secure scalable system that can work with on premises

The cloud can be made more secure than on premises

High speed redundant direct connect lines

Load billions of rows in minutes

All data in private VPC

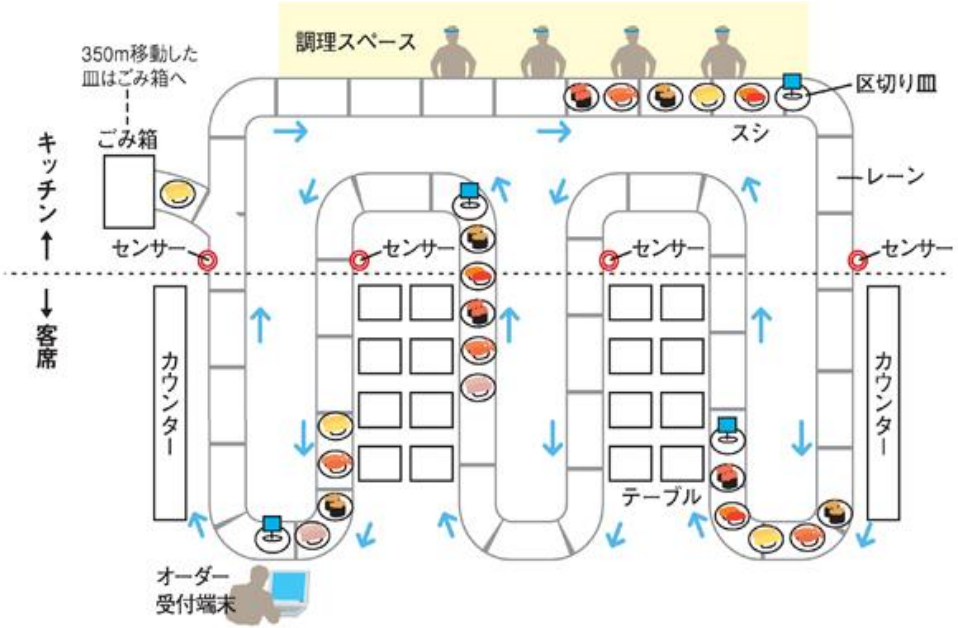
All data encrypted with private on-premises hardware keys

Encryption of data, transport, backups, partial spills

Audit of all SQL actions

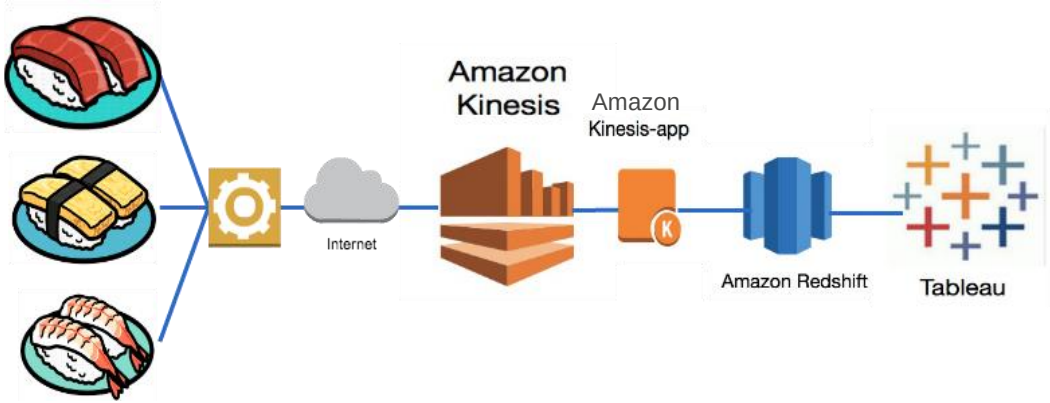
Audit of all configuration changes

Sushiro – Real-time streaming from IoT & analysis



Sushiro – Real-time streaming & analysis

Real-time data ingested by Amazon Kinesis is analyzed in Amazon Redshift



380 stores stream live data from Sushi plates

Inventory information combined with consumption information near real-time

Forecast demand by store, minimize food waste, and improve efficiencies

Data warehouses can support real-time data

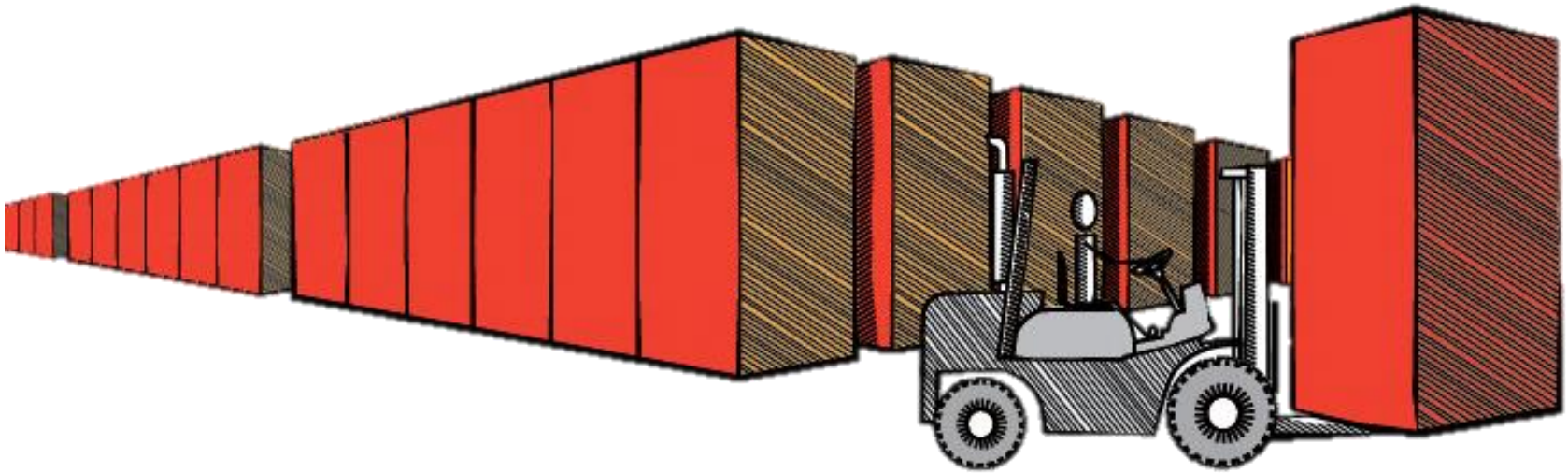
Big data does not mean batch

- Can be streamed in
- Can be processed in near real time
- Can be used to respond quickly to requests

You can mix and match

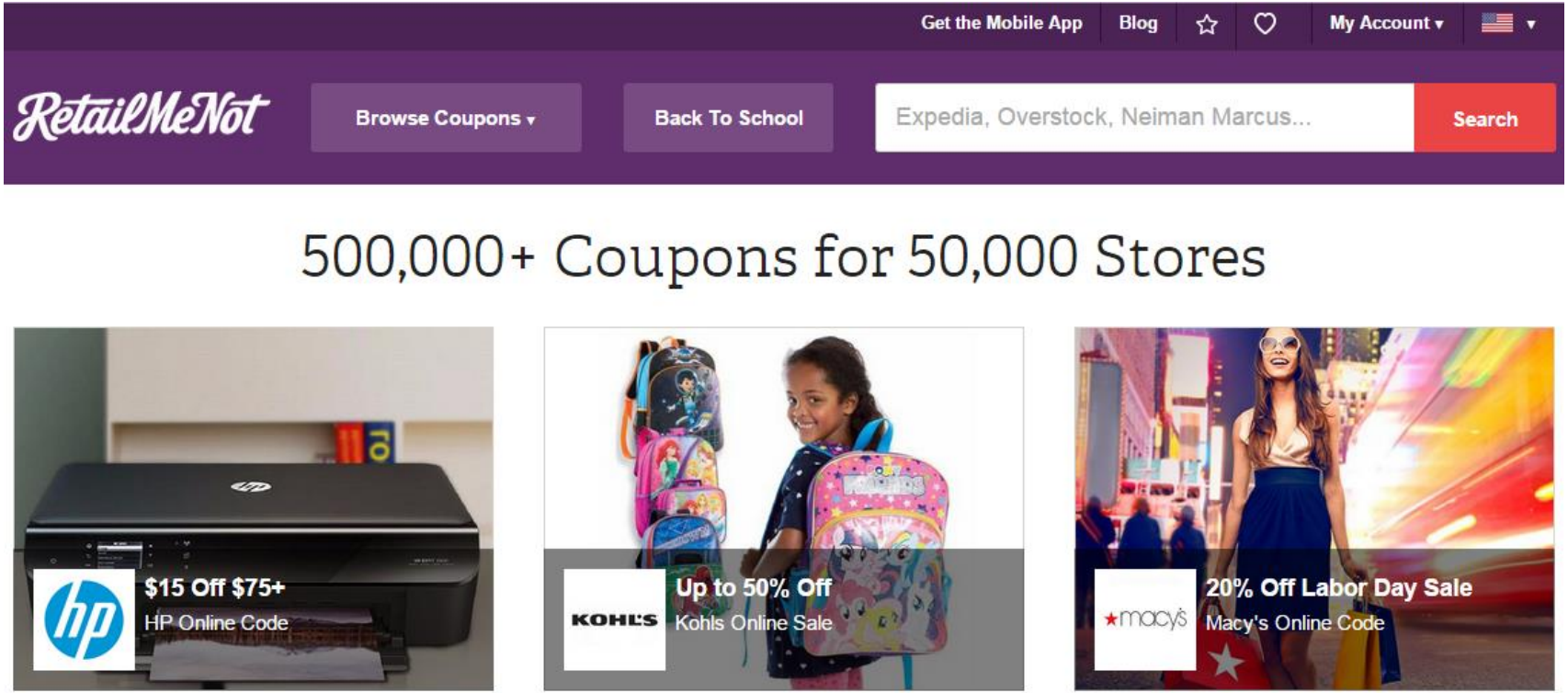
- On premises and cloud
- Custom development and managed services
- Infrastructure with managed scaling, security

In sum…



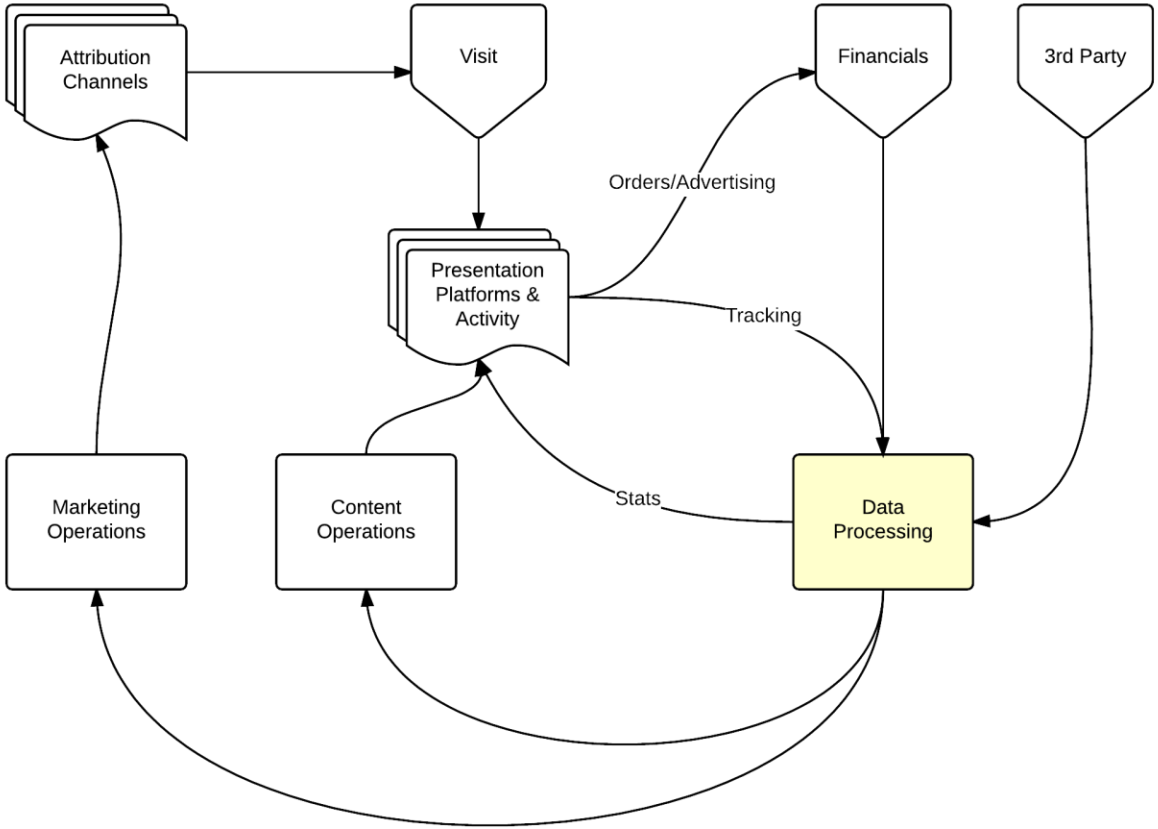
Amazon Redshift: Spend time with your data, not your database


RetailMeNot



RetailMeNot inc.

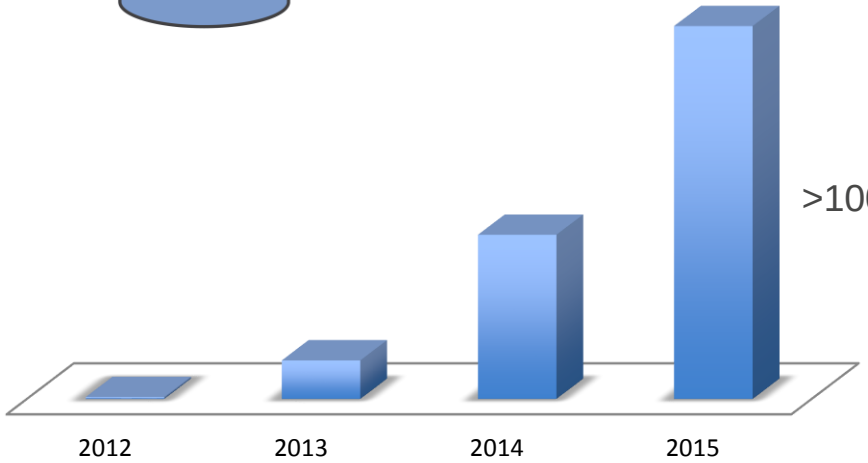
Our Data



RetailMeNot inc.

Our data

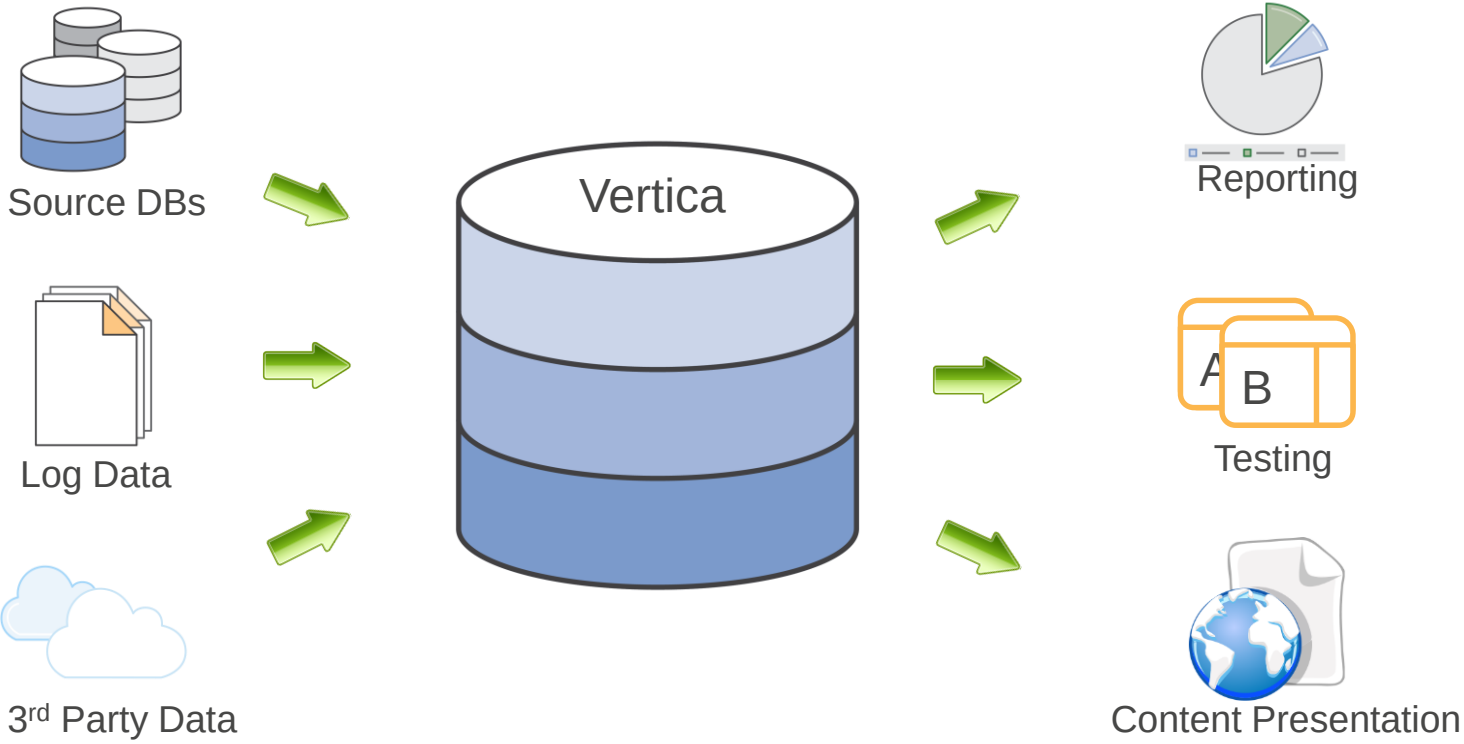
100s of TBs in Data Warehouses



>100% Year over Year Data Growth

RetailMeNot inc.

The legacy

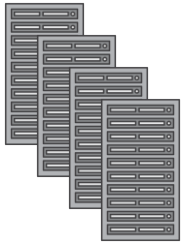


RetailMeNot inc.

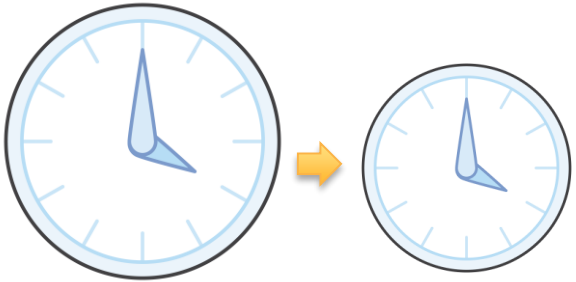
Pain points

Fire Fights

Query Traffic Jams



Scaling



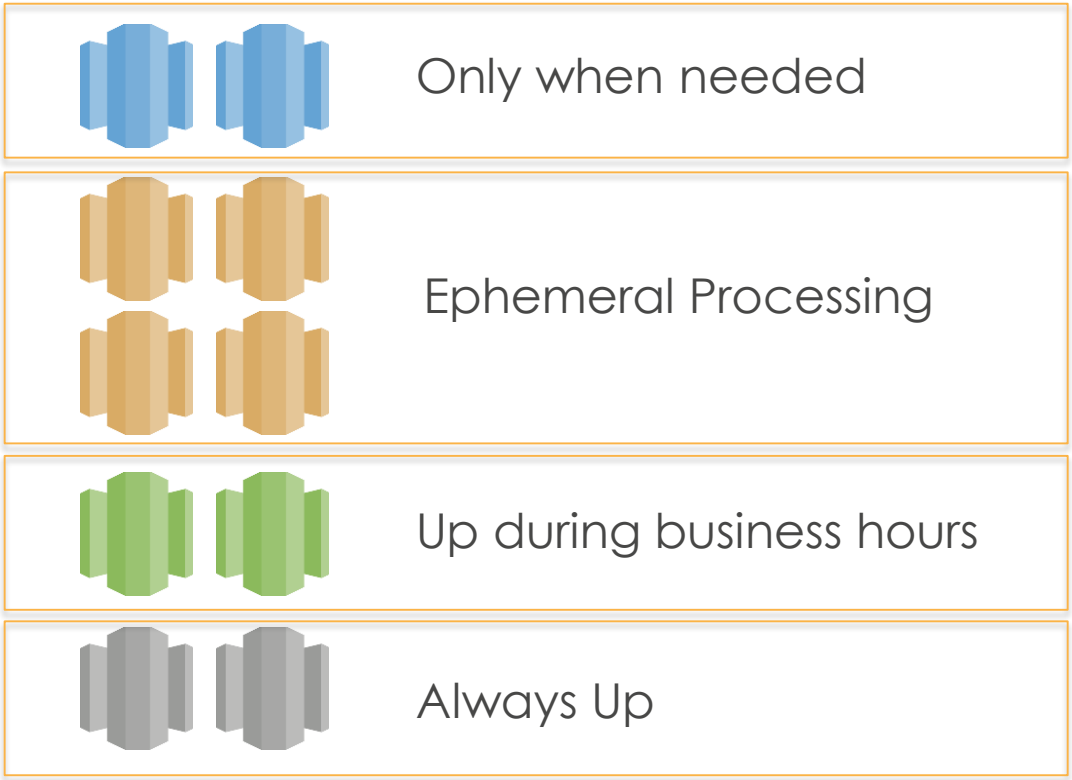
Processing Windows

RetailMeNot inc.

Adopting cloud strategies

RetailMeNot inc.

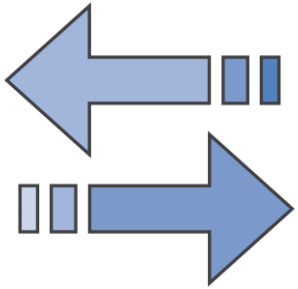
On-demand breakdown



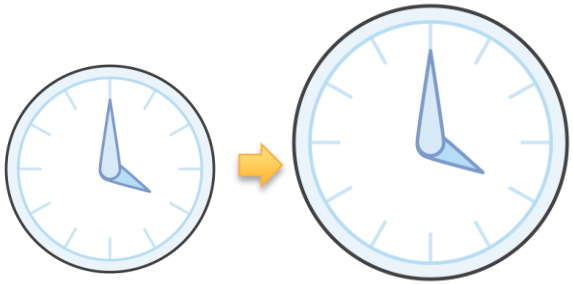
RetailMeNot inc.

Benefits to the data team

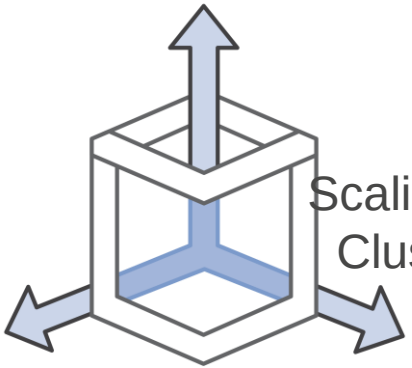
Fire Fights



Scaling Number of Clusters



Processing Windows

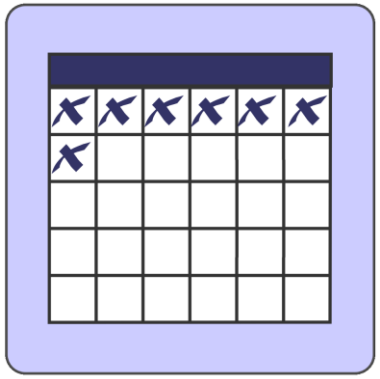


Scaling the Size of Clusters

RetailMeNot inc.

DOH!

Reserved Instances



Automated Cluster Shut Down

Sort/Distribution Keys
For Joins

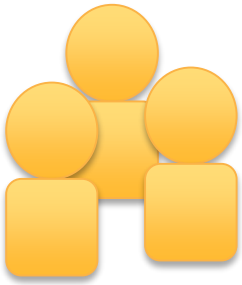
Automated vs. Manual Backups

RetailMeNot inc.

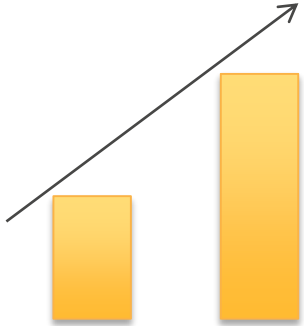
Benefits to the business

50% cost reduction for instances

$0 Licensing



50% Reduced time on administration



100% Growth of Internal Customers

RetailMeNot inc.

Q&A

Thank you!

Remember to complete your evaluations!

Related Sessions

Hear from other customers discussing their Amazon Redshift use cases:

- DAT308—How Yahoo! Analyzes Billions of Events with Amazon Redshift (Yahoo)
- ISM303—Migrating Your Enterprise Data Warehouse to Amazon Redshift (Boingo Wireless and Edmunds)
- ARC303—Pure Play Video OTT: A Microservices Architecture in the Cloud (Verizon)
- ARC305—Self-Service Cloud Services: How J&J Is Managing AWS at Scale for Enterprise Workloads
- BDT306—The Life of a Click: How Hearst Publishing Manages Clickstream Analytics with AWS
- DAT311—Large-Scale Genomic Analysis with Amazon Redshift (Human Longevity)
- BDT314—Running a Big Data and Analytics Application on Amazon EMR and Amazon Redshift with a Focus on Security (Nasdaq)
- BDT316—Offloading ETL to Amazon Elastic MapReduce (Amgen)
- BDT401—Amazon Redshift Deep Dive (TripAdvisor)