Overview of Amazon Redshift

Julien Simon, Principal Technical Evangelist, Amazon Web Services

julsimon@amazon.fr

@julsimon

Amazon Redshift

Fast, simple, petabyte-scale data warehousing for less than $1,000/TB/Year

amazon web services

Overview

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

Amazon Redshift works with your existing analysis tools

Amazon Redshift

JDBC/ODBC

actian
pentaho
ACTUATE The BIRT Company
periscope
birst
Qlik
ClearStory DATA Now You See It
SiSense
Chart.io Turn data to insight
+ableau SOFTWARE
Logi ANALYTICS
TIBCO
looker
Yellowfin
MicroStrategy
MODE

amazon web services

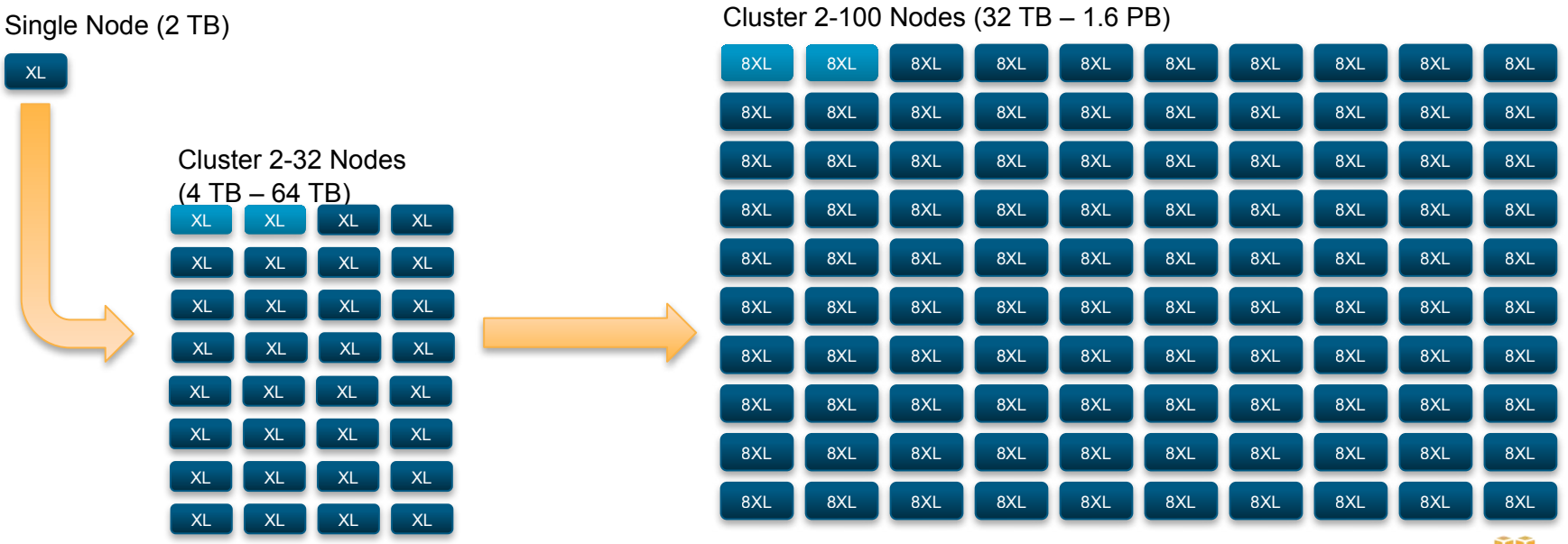
Amazon Redshift Scalability

Dense Storage Node (dw1.xlarge)
2 TB, 16 GB RAM, 2 cores

Dense Compute Node (dw2.large)
0.16 TB, 16 GB RAM, 2 cores

8XL Dense Storage Node (dw1.8xlarge)
16 TB, 128 GB RAM, 16 cores, 10 GigE

8XL Dense Compute Node (dw2.8xlarge)
2.56 TB, 128 GB RAM, 16 cores, 10 GigE



amazon web services

Amazon Redshift Value

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

- Node count x price per hour
- No charge for leader node
- No up-front costs
- Pay as you go

Our new Dense Storage (HDD) instance type

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

amazon web services

Architecture

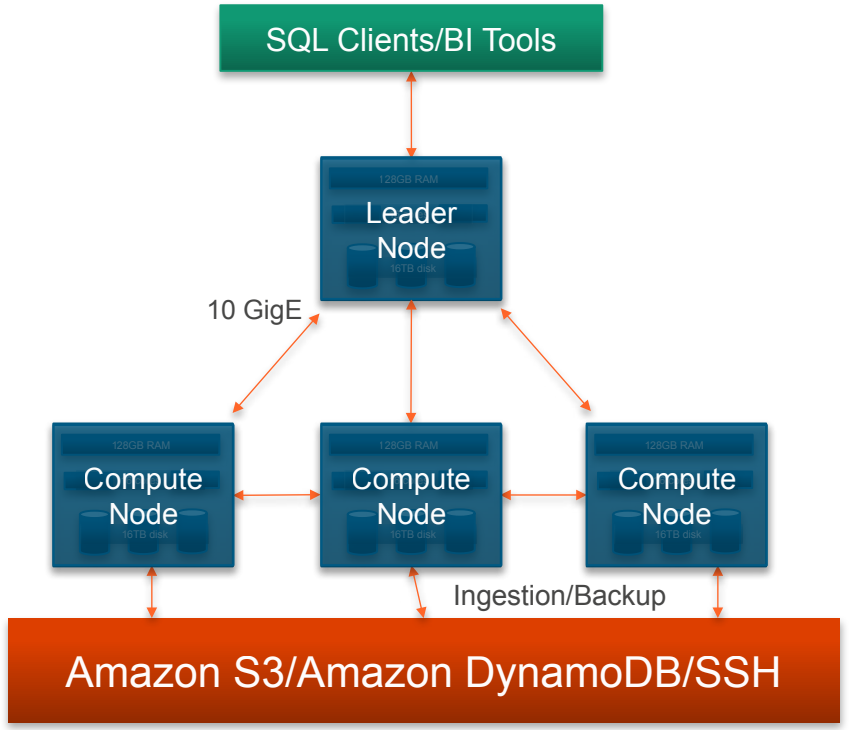
Amazon Redshift Architecture

Leader Node

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

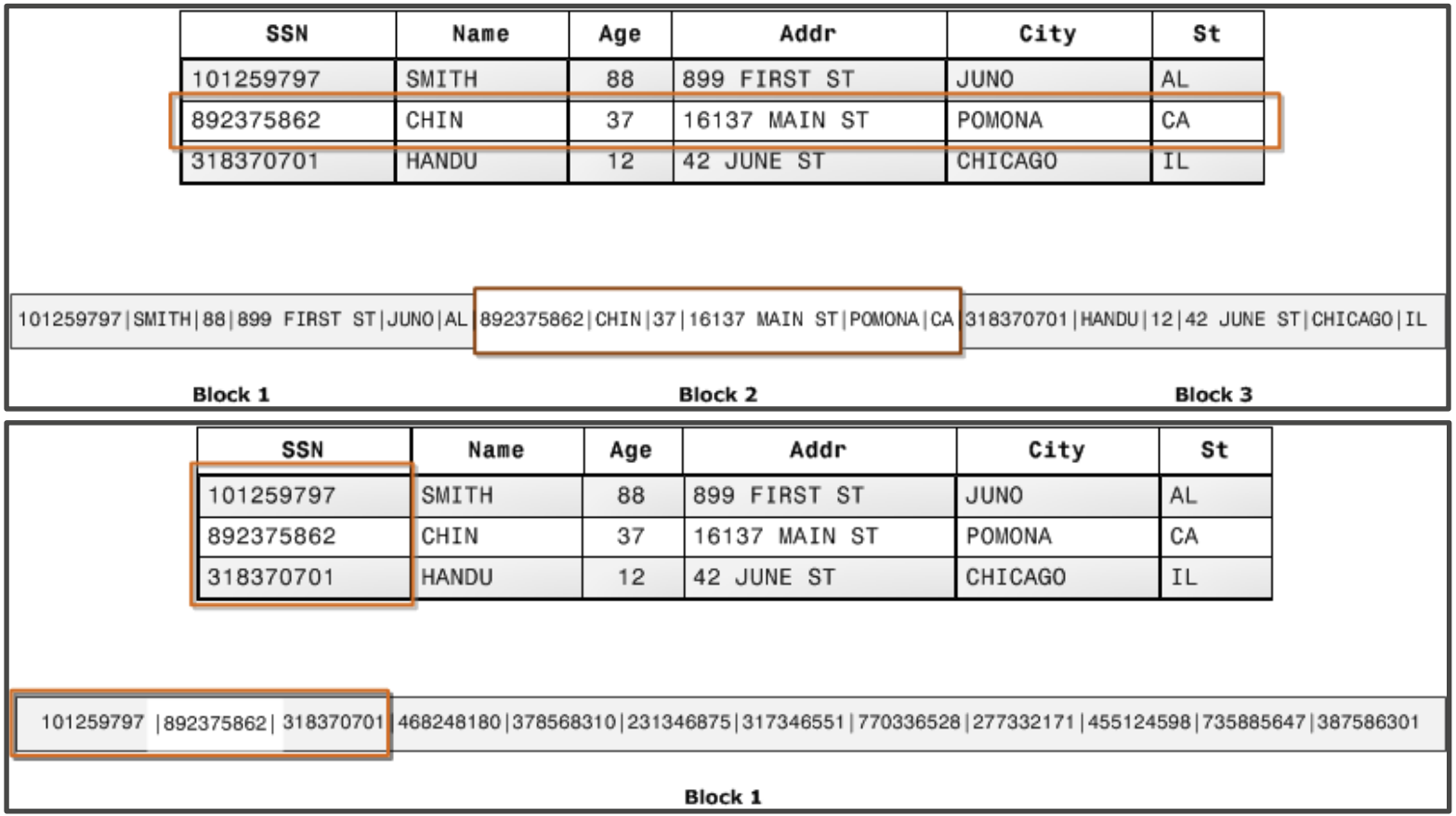
Compute Nodes

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



amazon web services

Row storage vs columnar storage



https://docs.aws.amazon.com/fr_fr/redshift/latest/dg/c_columnar_storage_disk_mem_mgmnt.html

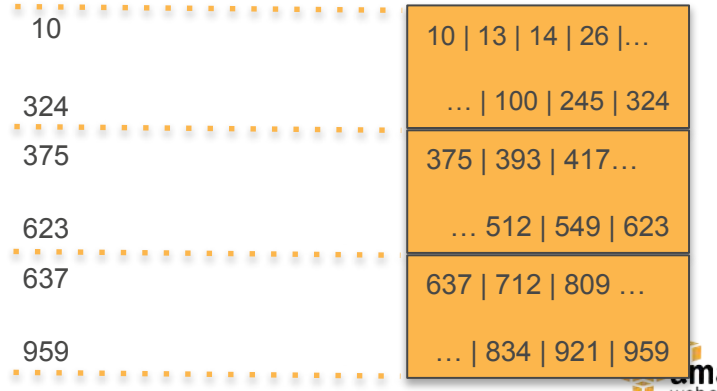
Amazon Redshift Performance

Dramatically reduce I/O

- Columnar 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
```

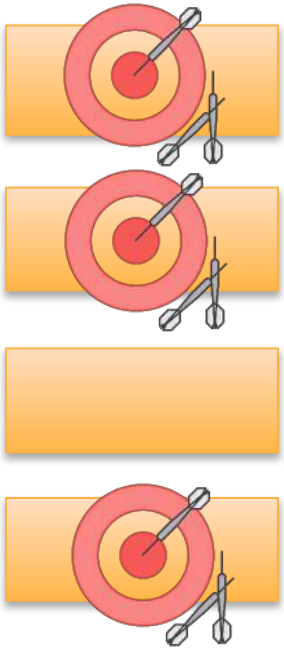


Amazon Redshift Performance

Sort Keys and Zone Maps

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

Unsorted Table



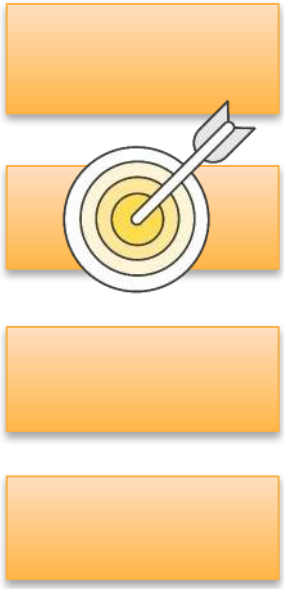
MIN: 01-JUNE-2013
MAX: 20-JUNE-2013

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

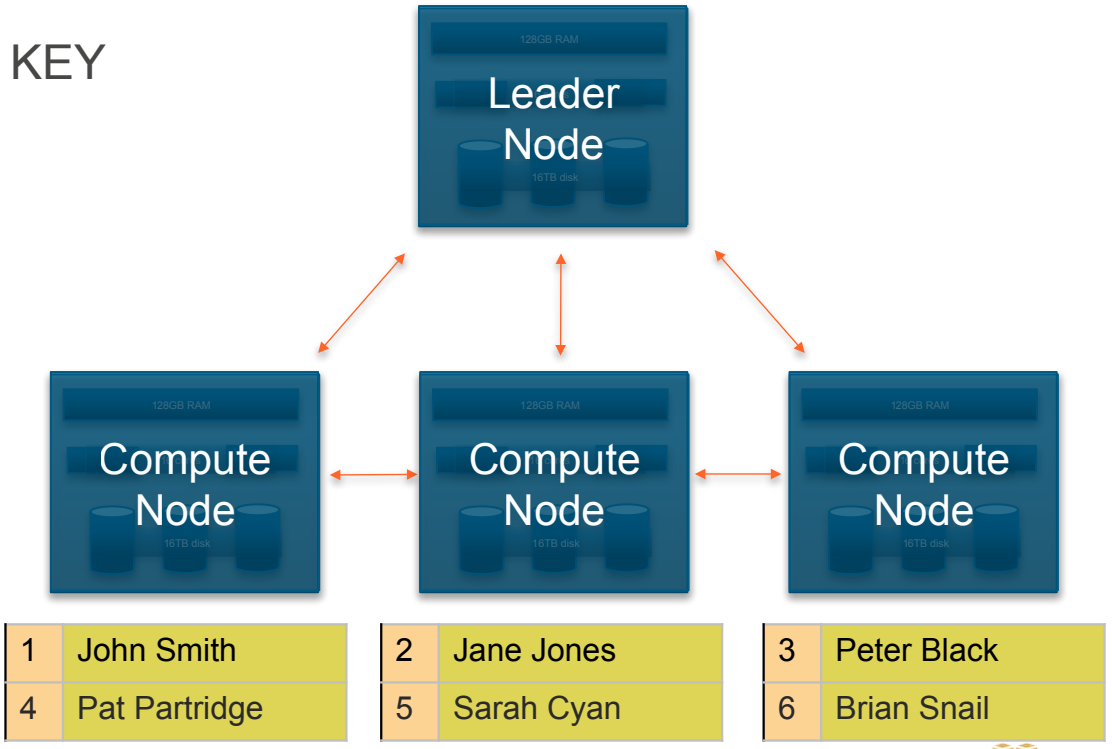
MIN: 19-JUNE-2013
MAX: 24-JUNE-2013

Amazon Redshift Performance

Distribution Keys

3 policies: EVEN (default), ALL, KEY

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



amazon web services

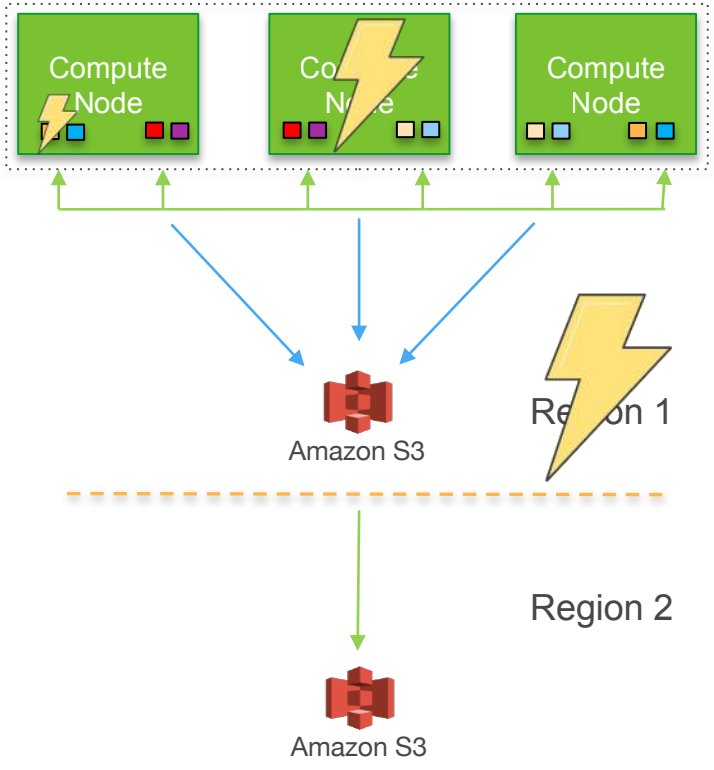
Amazon Redshift Robustness

Data availability

- Multiple copies within cluster
- Continuous and incremental backups to Amazon S3

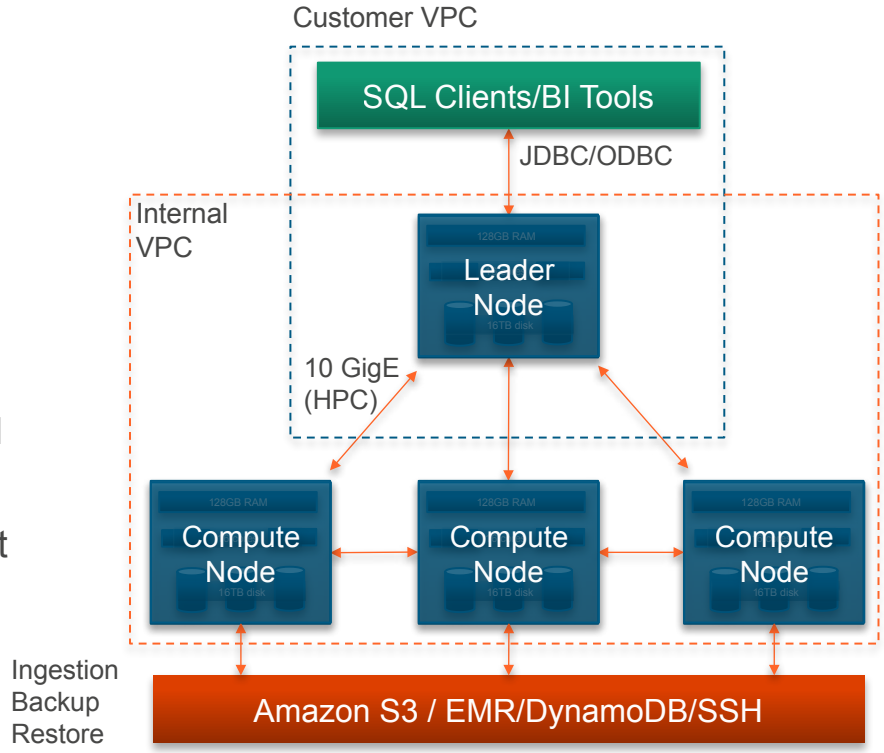
Fault tolerance

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



Amazon Redshift Security

- Load encrypted from S3
- SSL to secure data in transit
- 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



amazon web services

Case Studies

Photobox

http://www.lemagit.fr/etude/Photobox-consolide-et-analyse-ses-donnees-avec-AWS-RedShift

Maxime Mezin, Data & Photo Science Director:

"*L'entrepôt de données ne comportait que les données du site e-commerce liées aux ventes. Alors que nous avions la volonté d'intégrer des données du service clients et des données d'analyse (…) Nous avions atteint la limite du stockage de la base existante, et cela ne marchait pas très bien en termes de performances*"

"*Avec Redshift, la rapidité d'exécution des traitements a été multipliée par 10. Sans parler de la vitesse de chargement des données*"

- 2 Redshift clusters : 1 for historical data, 1 for real-time processing (SSD)
- Total Cost of Ownership divided by 7 (90K€→13K€)

Financial Times

https://aws.amazon.com/solutions/case-studies/financial-times/

- BI analysis of reader traffic, in order to decide which stories to cover
- Conventional data warehouse running on old-guard technology
- Scalability issues, impossible to perform real-time analytics → Amazon Redshift PoC
- Amazon Redshift performed so quickly that some analysts thought it was malfunctioning ☺

John O'Donovan, CTO: "*Amazon Redshift is the single source of truth for our user data.*"

"*Some of the queries we're running are 98 percent faster, and most things are running 90 percent faster (…) and the ability to try Redshift out before having to invest a significant amount of capital was a huge bonus*."

"*Being able to explore near-real-time data improves our decision making massively. We can make decisions based on what's happening now rather than what happened three or four days ago.*"

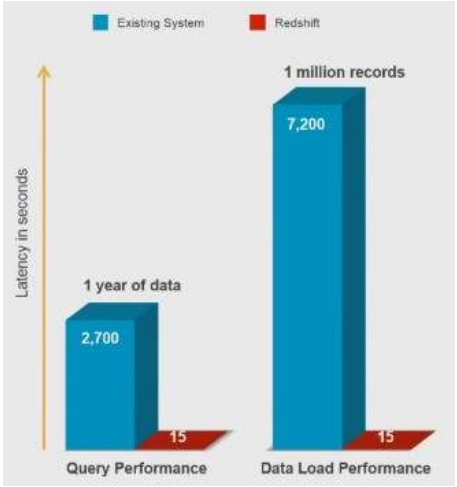
- Total Cost of Ownership divided by 4

Boingo

https://www.youtube.com/watch?v=58URZbp1voY

- Largest operator of airport wireless hotspots in the world: 1M+ hotspots,100+ countries
- About 15 TB of data, growing at 2-3 TB per year
- Legacy platform: low performance, heavy admin, high cost
- Evaluated Amazon Redshift and two other vendors
- Selected Amazon Redshift and migrated in 2 months

6-7x less expensive than alternatives



Queries
180x faster

Data load
480x faster

amazon web services

Amazon Redshift

Fast, simple, petabyte-scale data warehousing for less than $1,000/TB/Year

amazon web services

Additional Resources

Resources

Detail pages

- http://aws.amazon.com/redshift
- https://aws.amazon.com/marketplace/redshift/

Best practices

- http://docs.aws.amazon.com/redshift/latest/dg/c_loading-data-best-practices.html
- http://docs.aws.amazon.com/redshift/latest/dg/c_designing-tables-best-practices.html
- http://docs.aws.amazon.com/redshift/latest/dg/c-optimizing-query-performance.html

Contact

julsimon@amazon.fr
@julsimon

amazon web services