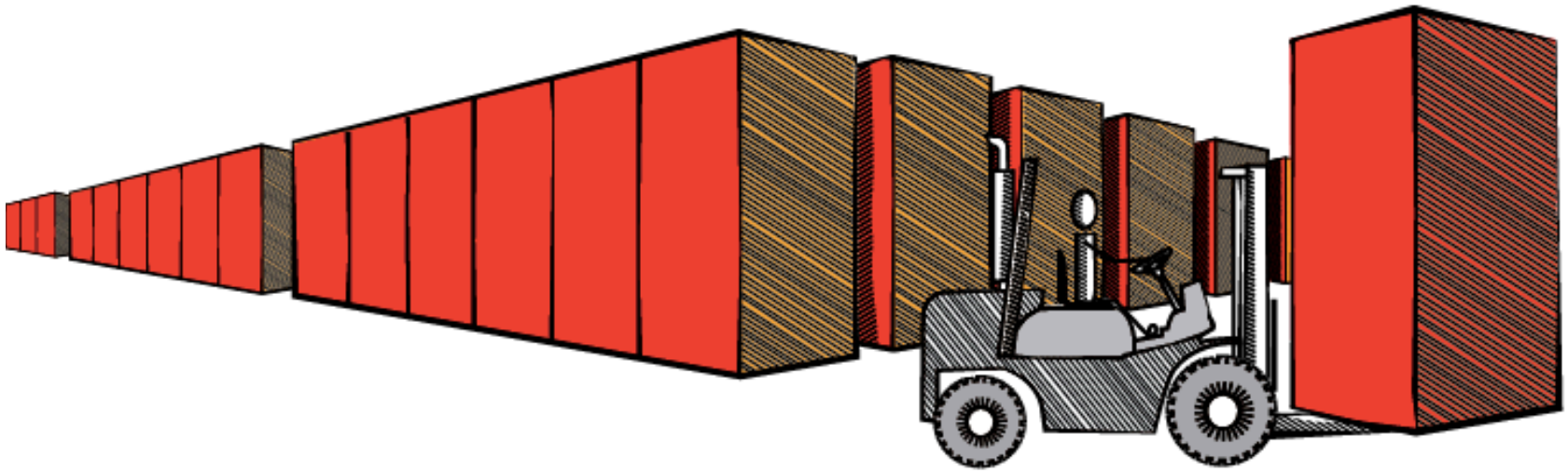


Amazon Redshift

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

Agenda

9:00-9:45 – Introduction to Redshift
9:45-10:15 – RedShift Security
10:15-10:45 – RedShift Operations
10:45-11:15 – Redshift Table & Schema Design Loading
11:15-12:00 – Redshift Table & Schema Design Loading Lab

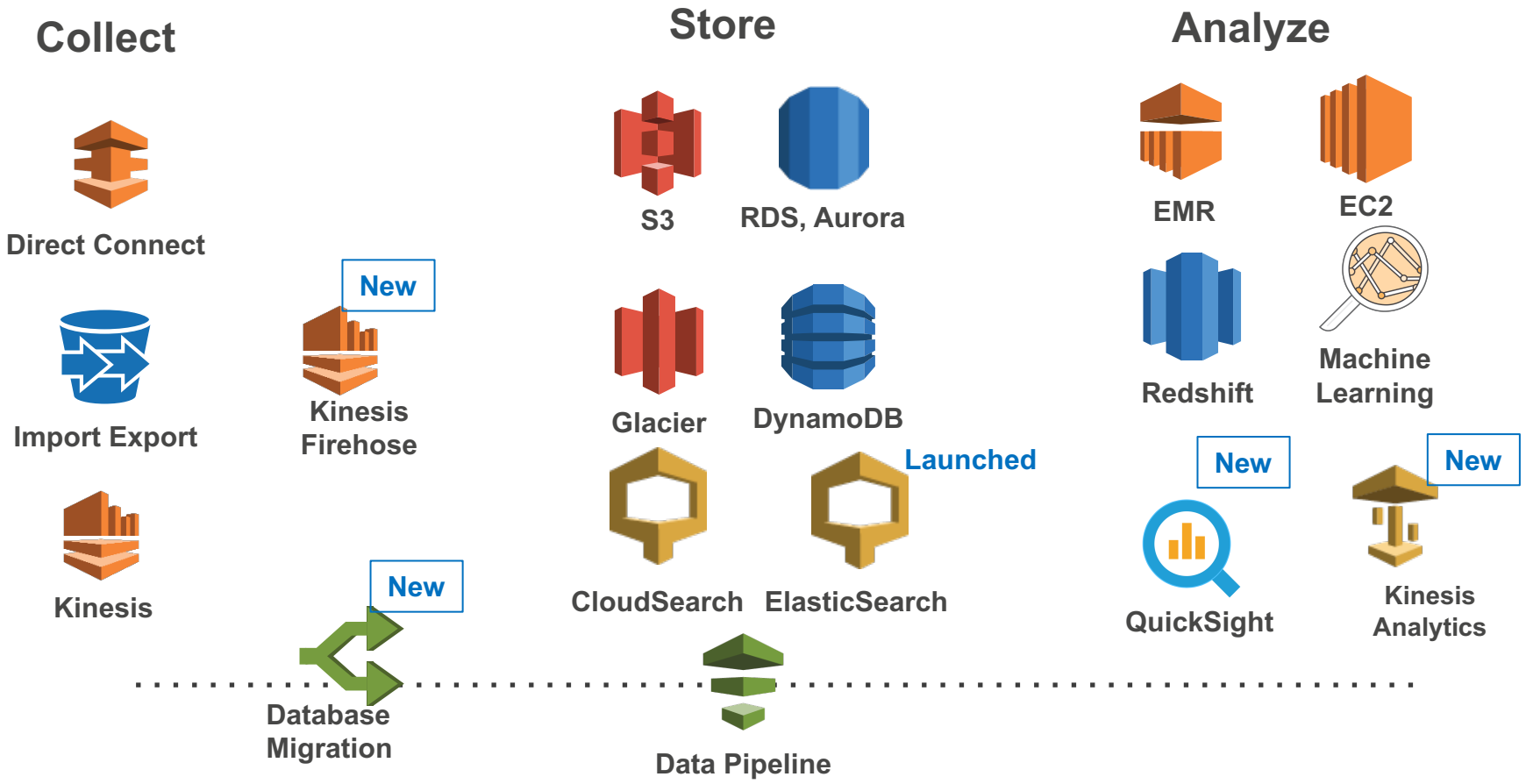
12:00-1:00 – Lunch

1:00 – 1:45 – RedShift Data Loading
1:45– 2:45 – Data Loading Lab
2:45-3:00 – Design Considerations
3:00 – Next Steps

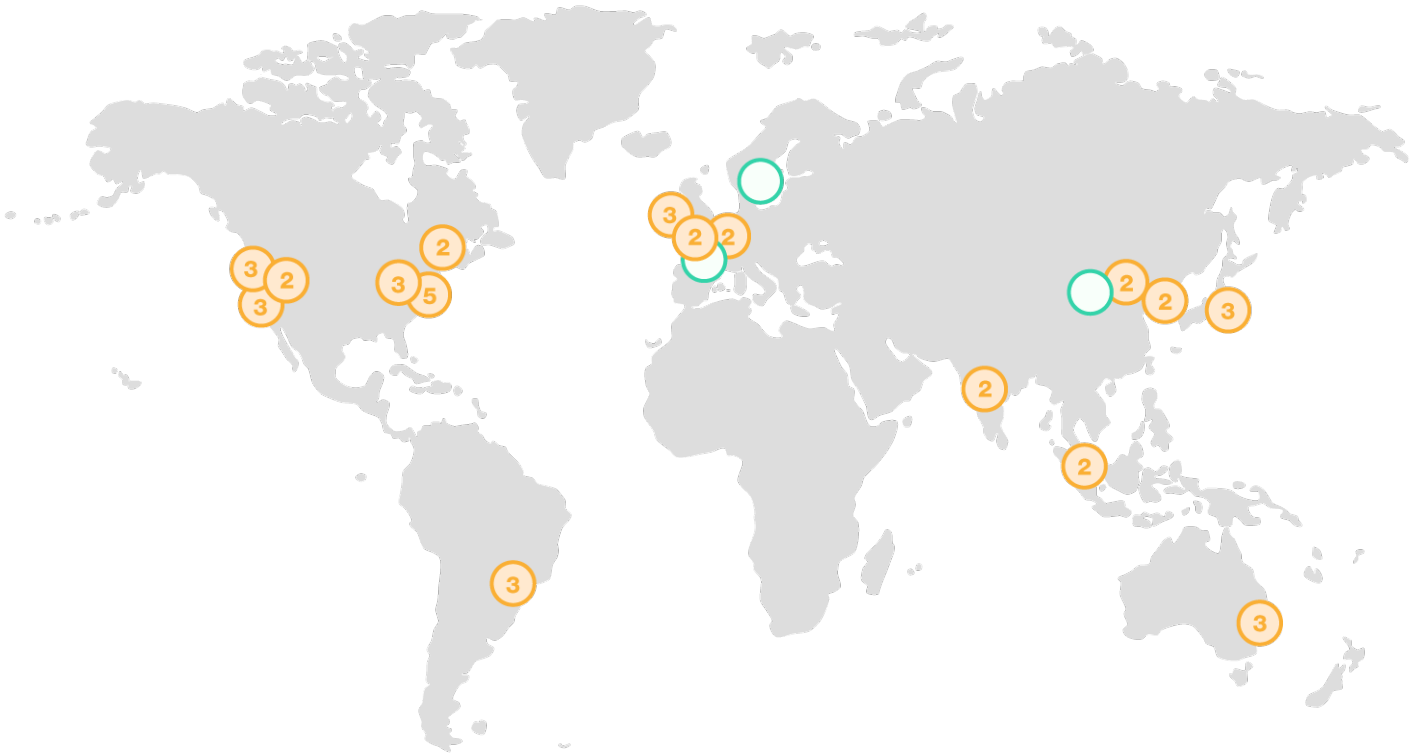
We start with the customer… and innovate

Customers told us…	We created…
Managing databases is painful & difficult	✓ Amazon RDS
SQL DBs do not perform well at scale	✓ Amazon DynamoDB
Hadoop is difficult to deploy and manage	✓ Amazon EMR
DWs are complex, costly, and slow	✓ Amazon Redshift
Commercial DBs are punitive & expensive	✓ Amazon Aurora
Streaming data Is difficult to capture & analyze	✓ Amazon Kinesis
BI Tools are expensive and hard to manage	✓ Amazon QuickSight

AWS Big Data Portfolio



Global Footprint



Redshift

16 Regions; 16 Availability Zones; 76 Edge Locations

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

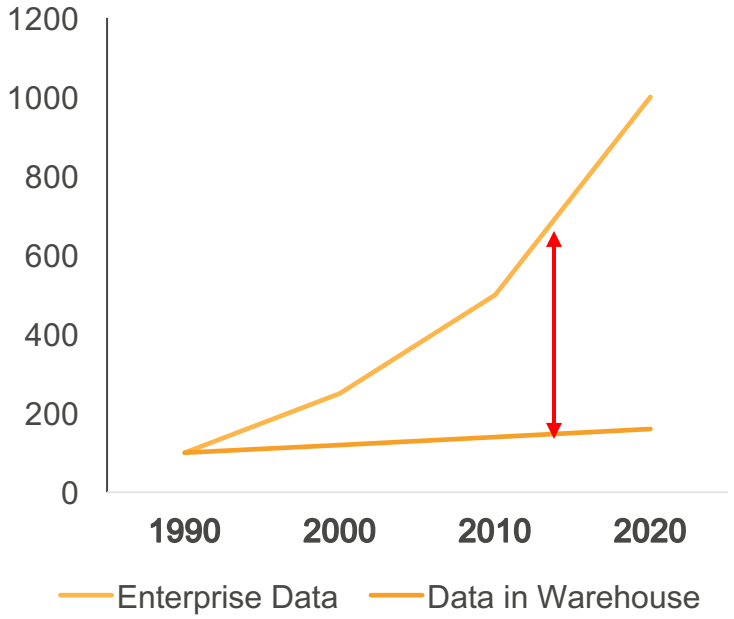
Sell to central IT

Multi-year commitment

Multi-year deployments

Multi-million dollar deals

… Leads to dark data



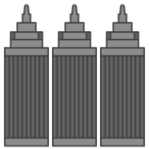
This is a narrow view

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

NTT Docomo | Telecom

FINRA | Financial Svcs

Philips | Healthcare

Yelp | Technology

NASDAQ | Financial Svcs

The Weather Company | Media

Nokia | Telecom

Pinterest | Technology

Foursquare | Technology

Coursera | Education

Coinbase | Bitcoin

Amazon | E-Commerce

Etix | Entertainment

Spuul | Entertainment

Vivaki | Ad Tech

Z2 | Gaming

Neustar | Ad Tech

SoundCloud | Technology

BeachMint | E-Commerce

Civis | Technology

Amazon Redshift Architecture

Leader Node

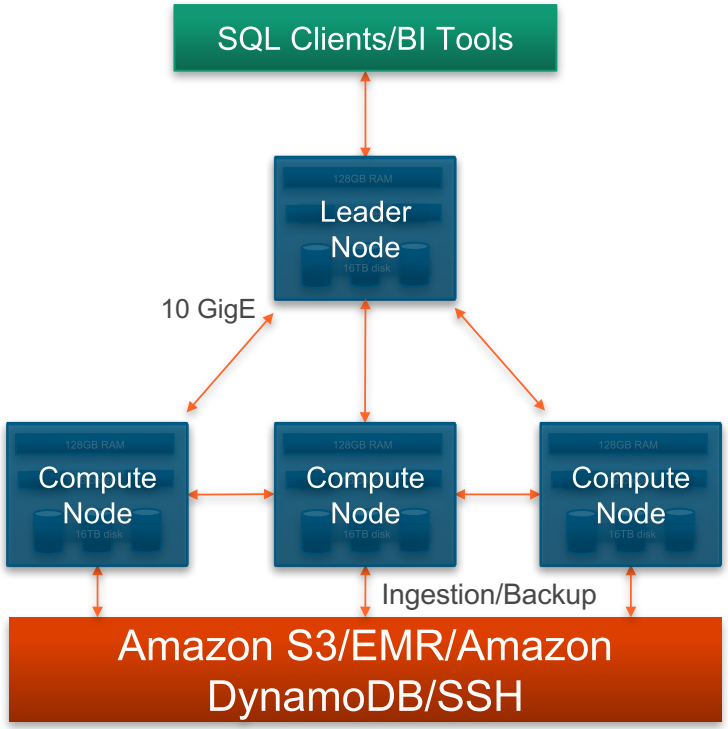
- Simple SQL end point
- Stores metadata
- Optimizes query plan
- Coordinates query execution

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
- DS1/DS2: HDD; scale from 2 TB to 2 PB



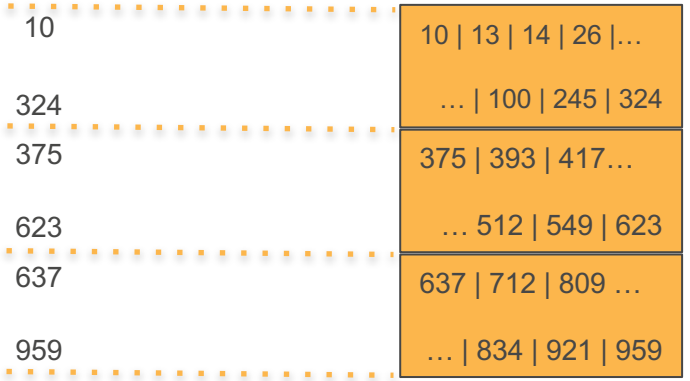
Benefit #1: Amazon Redshift is fast

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
```

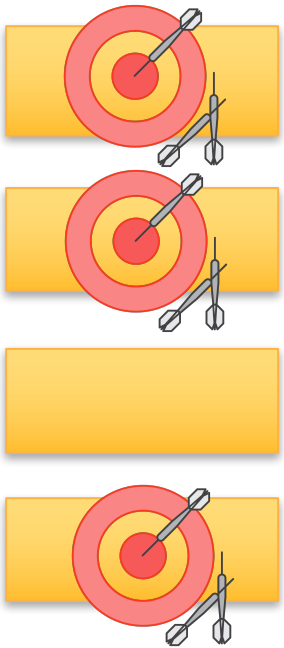


Benefit #1: Amazon Redshift is fast

Sort Keys and Zone Maps

SELECT COUNT(*) FROM LOGS WHERE DATE = '09-JUNE-2016'

Unsorted Table



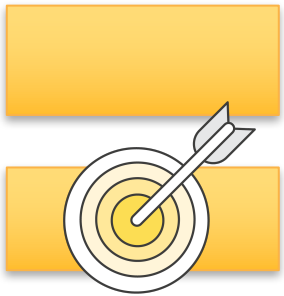
MIN: 01-JUNE-2016
MAX: 20-JUNE-2016

MIN: 08-JUNE-2016
MAX: 30-JUNE-2016

MIN: 12-JUNE-2016
MAX: 20-JUNE-2016

MIN: 02-JUNE-2016
MAX: 25-JUNE-2016

Sorted By Date



MIN: 01-JUNE-2016
MAX: 06-JUNE-2016

MIN: 07-JUNE-2016
MAX: 12-JUNE-2016

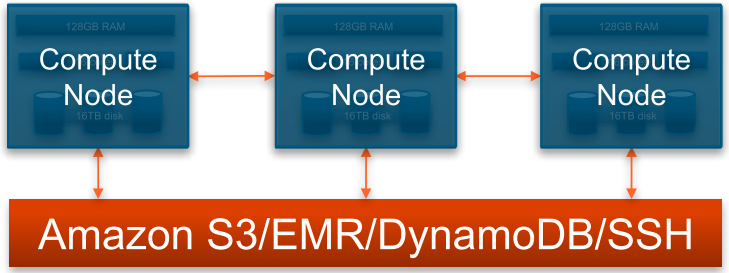
MIN: 13-JUNE-2016
MAX: 18-JUNE-2016

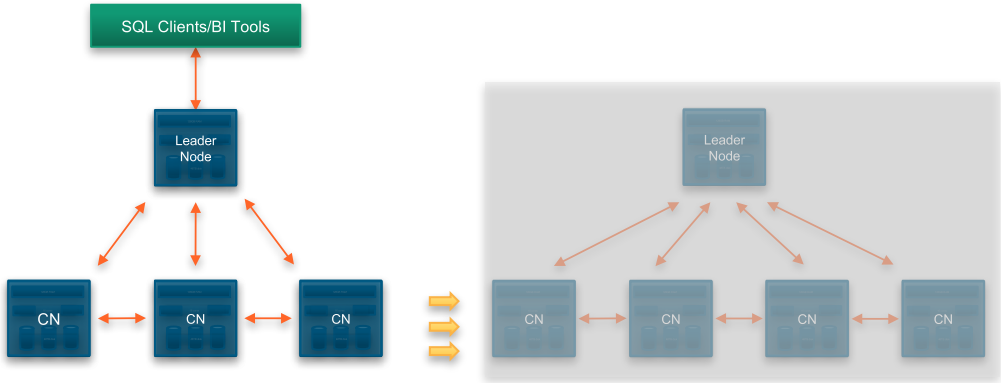
MIN: 19-JUNE-2016
MAX: 24-JUNE-2016

Benefit #1: Amazon Redshift is fast

Parallel and Distributed

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

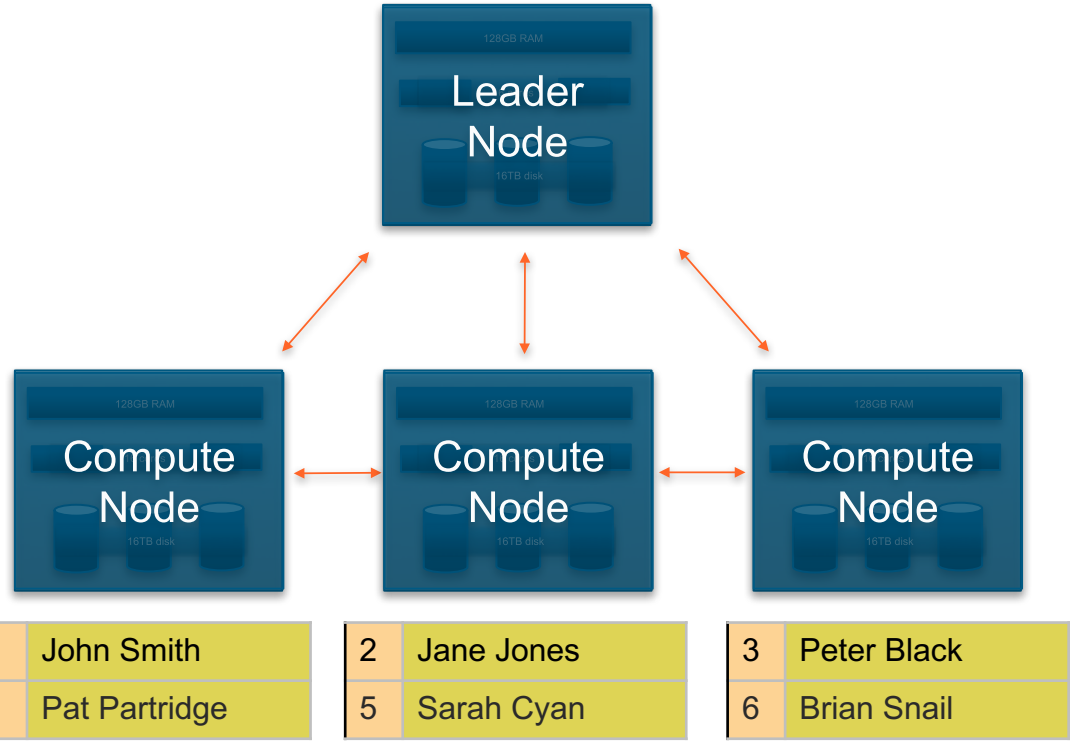




Benefit #1: Amazon Redshift is fast

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

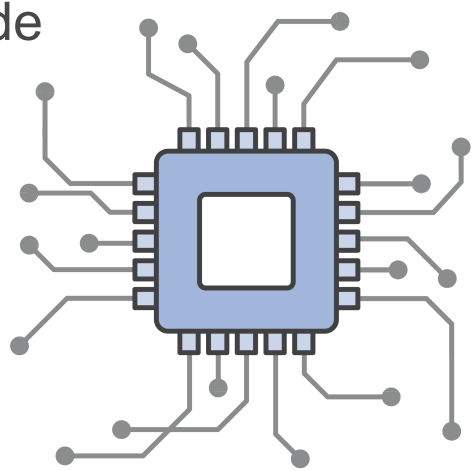
Enhanced networking, over 1M packets/sec/node

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 DS2.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 DC1.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 #2: Amazon Redshift is inexpensive

Amazon Redshift lets you start small and grow big

Dense Storage (DS2.XL)

2 TB HDD, 31 GB RAM, 2 slices/4 cores

Single Node (2 TB)

Cluster 2-32 Nodes (4 TB – 64 TB)

Note: Nodes not to scale

Dense Storage (DS2.8XL)

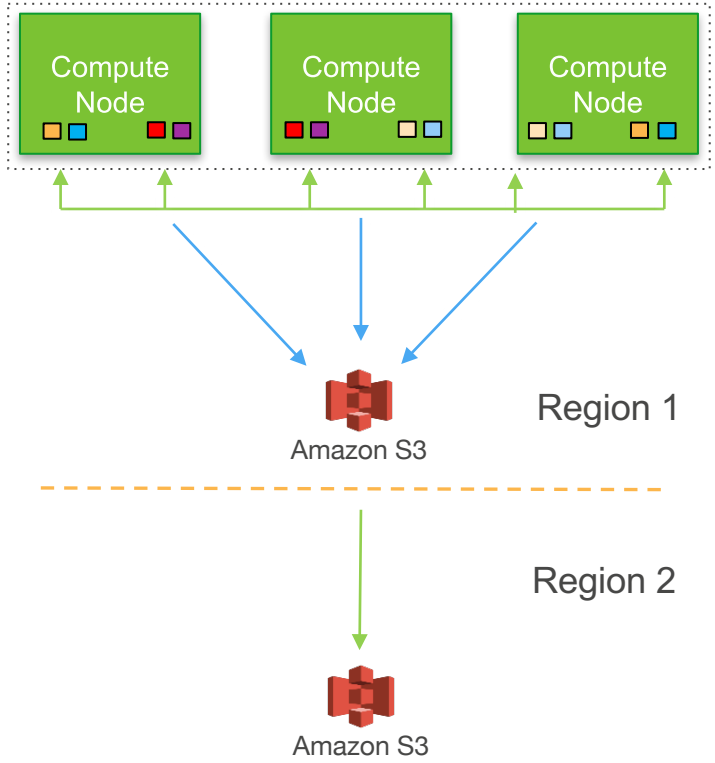
16 TB HDD, 244 GB RAM, 16 slices/36 cores, 10 GigE

Cluster 2-128 Nodes (32 TB – 2 PB)

Benefit #3: Amazon Redshift is fully managed

Continuous/incremental backups

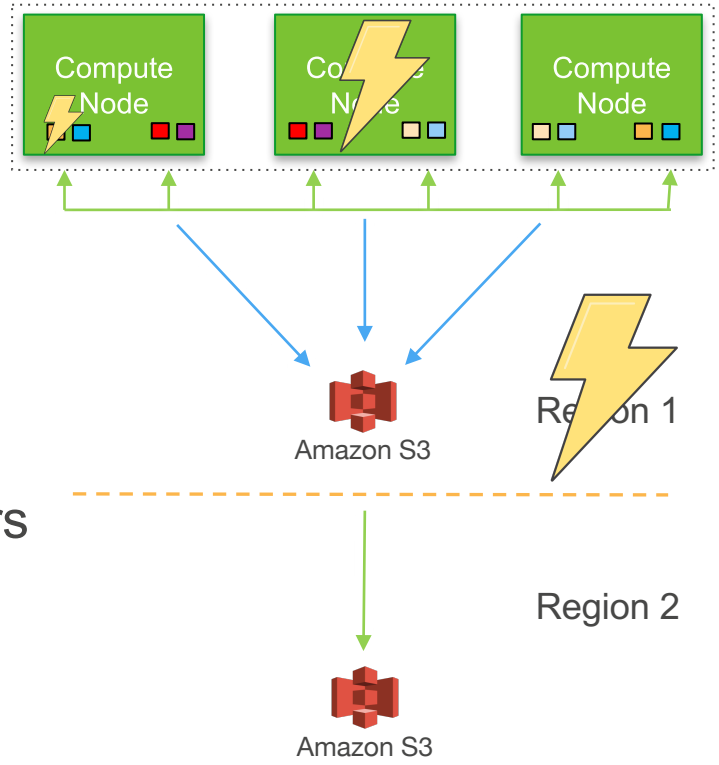
- Multiple copies within cluster
- Continuous and incremental backups to S3
- Continuous and incremental backups across regions
- Streaming restore



Benefit #3: Amazon Redshift is fully managed

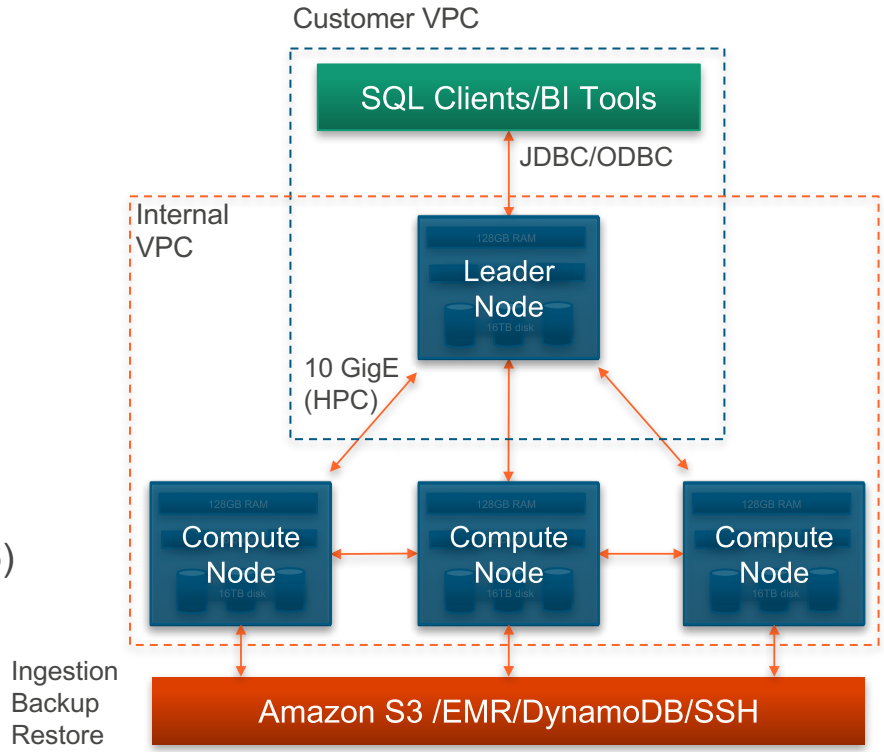
Fault tolerance

- 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 & KMS support
- Audit logging and AWS CloudTrail integration
- SOC 1/2/3, PCI-DSS, FedRAMP, BAA



Benefit #5: We innovate quickly

100+ new features added since launch
Release every two weeks
Automatic patching

2013

- Initial Release in US East (N. Virginia; US West (Oregon), EU (Ireland); Asia Pacific (Tokyo, Singapore, Sydney) Regions
- MANIFEST option for the COPY & UNLOAD commands
- SQL Functions: Most recent queries
- Resource-level IAM, CRC32
- Data Pipeline
- Event notifications, encryption, key rotation, audit logging, on-premises or AWS CloudHSM; PCI, SOC 1/2/3
- Cross-Region Snapshot Copy
- Audit features, cursor support, 500 concurrent client connections
- EIP Address for VPC Cluster

2014

- System Tables for query Tuning
- Dense Compute Nodes
- Gzip & Lzop; JSON , RegEx, Cursors
- EMR Data Loading & Bootstrap Action with COPY command; WLM concurrency limit to 50; support for the ECDH cipher suites for SSL connections; FedRAMP
- Cross-region ingestion
- Free trials & price reductions in Asia Pacific
- CloudWatch Alarm for Disk Usage
- AES 128-bit encryption; UTF-16; KMS Integration
- EU (Frankfurt); GovCloud Regions
- S3 Servier-side encryption support for UNLOAD
- Tagging Support for Cost-allocation

2015

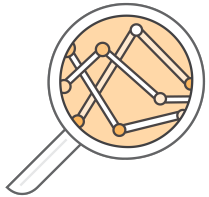
- New system views to tune table design and track WLM query queues
- Custom ODBC/JDBC drivers; Query Visualization
- Mobile Analytics auto export
- KMS for GovCloud Region; HIPAA BAA
- Interleaved sort keys
- New Dense Storage Nodes (DS2) with better RAM and CPU.
- New Reserved Storage Nodes: No, Partial & All Upfront Options
- Cross-region backups for KMS encrypted clusters
- Scaler UDFs in Python
- AVRO Ingestion; Kinesis Firehose; Database Migration Service (DMS)
- Modify Cluster Dynamically
- Tag-based permissions and BZIP2

2016

- WLM Queue-Hopping for timed-out queries
- Append rows & Export to BZIP-2
- Lambda for Clusters in VPC; Data Schema Conversion Support from ML Console
- US West (N. California) Region.

Benefit #6: Amazon Redshift is powerful

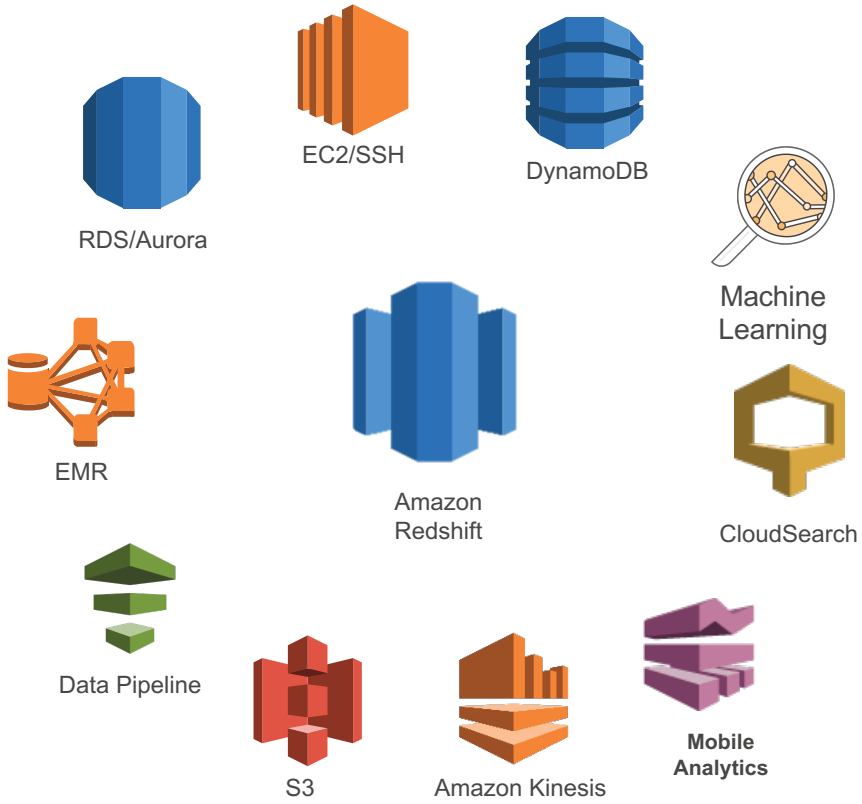
- Approximate functions
- User defined functions
- Machine Learning
- Data Science



Amazon ML

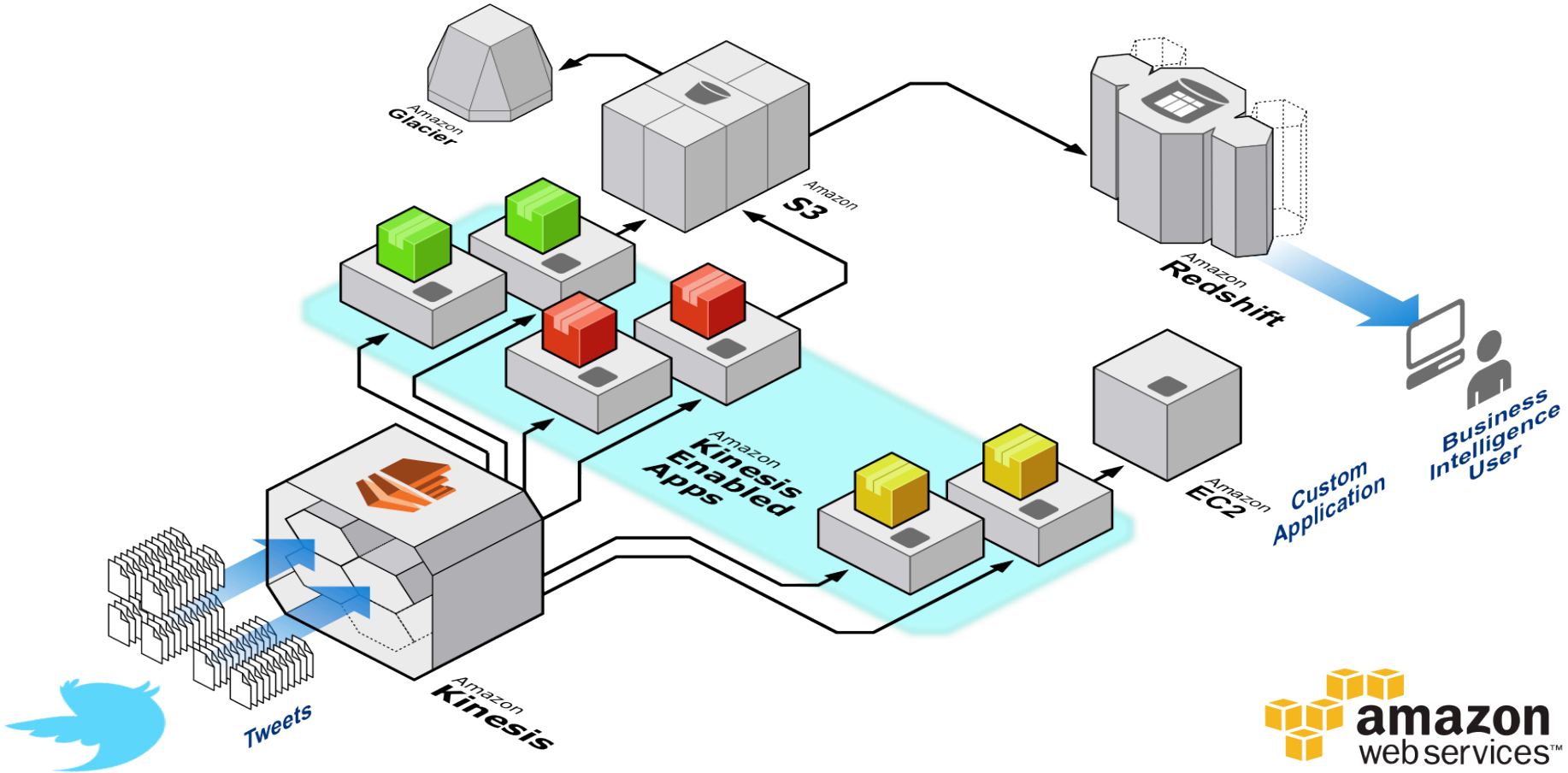
Benefit #7: Amazon Redshift has a large ecosystem

Benefit #8: Service oriented architecture

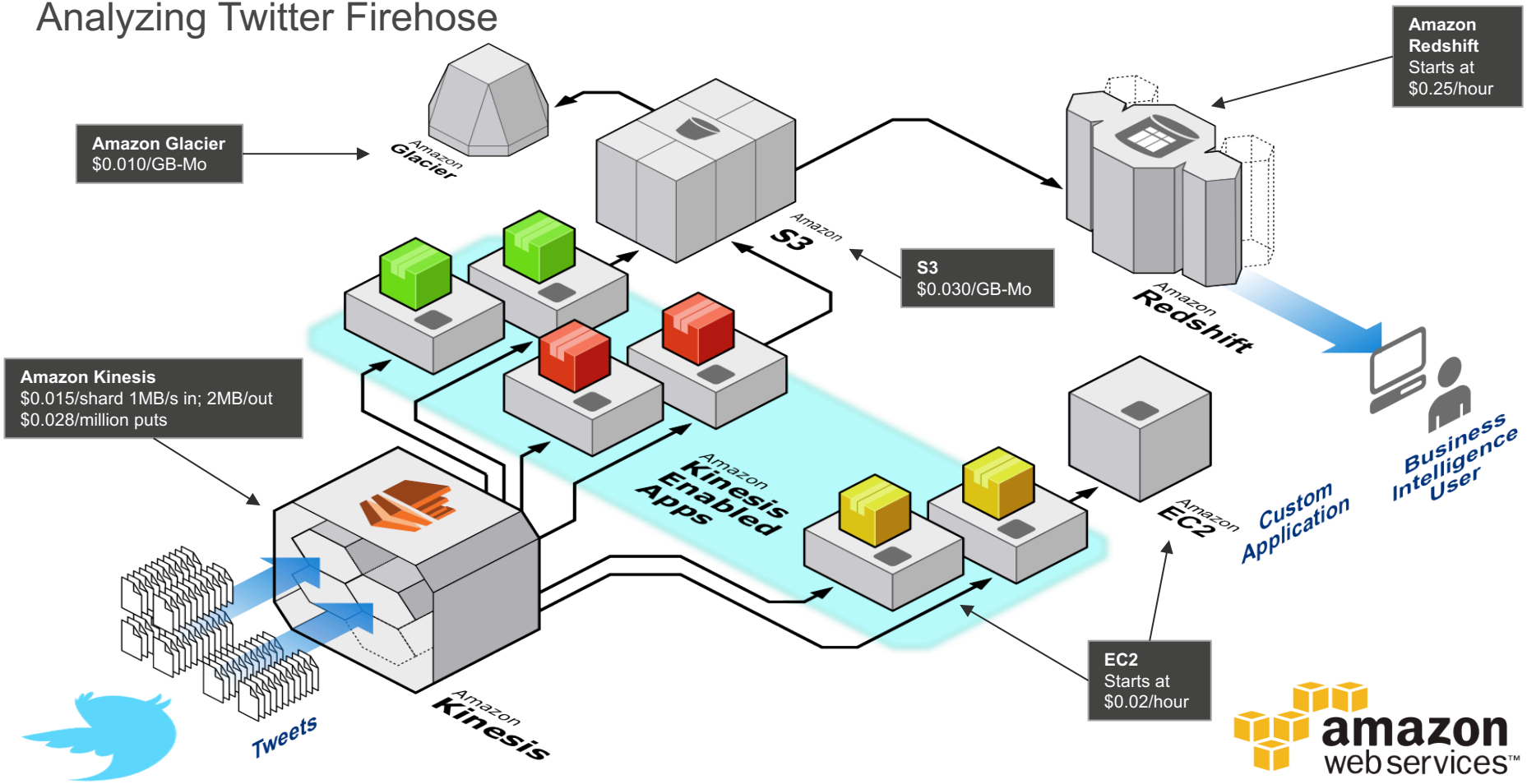


Use cases

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

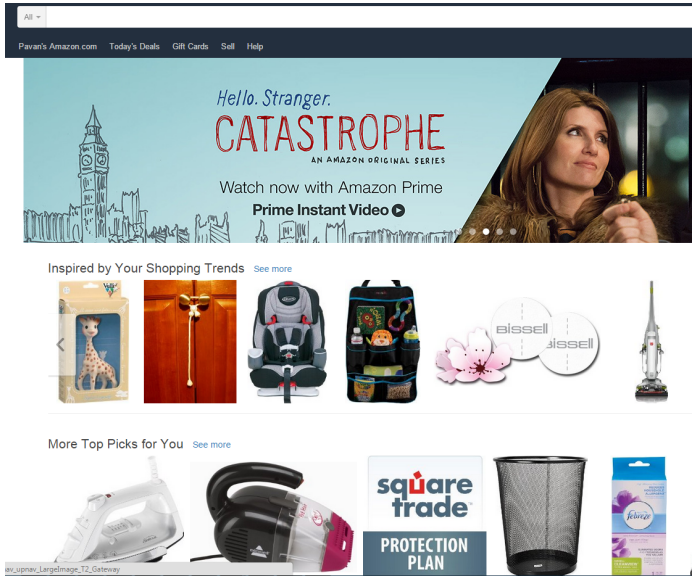
Web log analysis for Amazon.com

- 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.



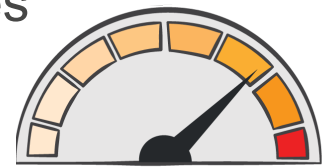
Data warehouses can be

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

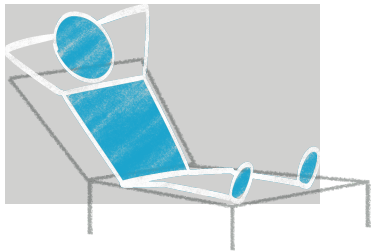
Load pipeline: 90 hours (Oracle) to 8 hours

64 clusters

800 total nodes

13PB provisioned storage

2 DBAs



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

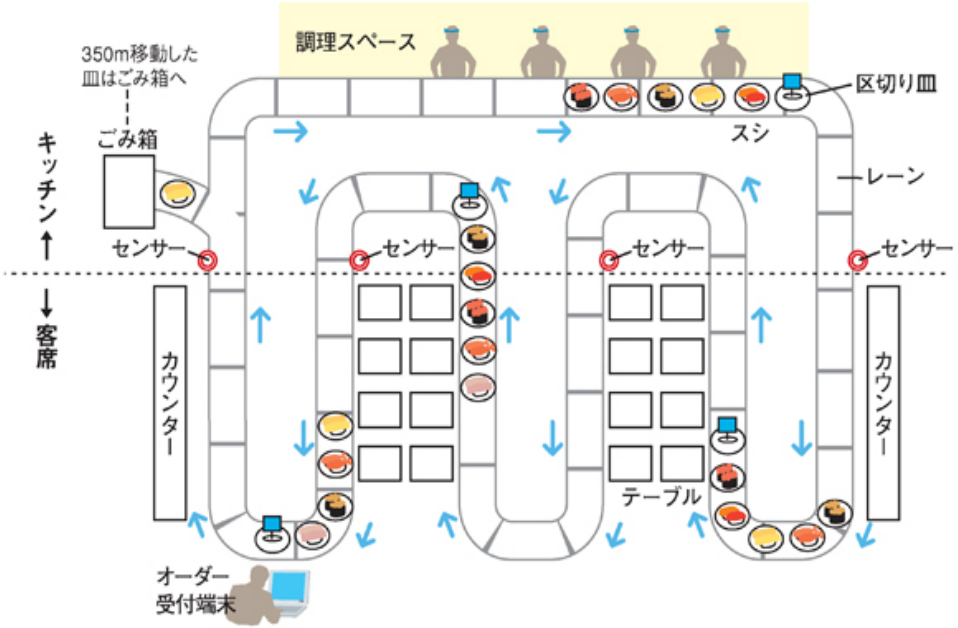
All data encrypted with private on-premises hardware keys

Encryption of data, transport, backups, partial spills

Audit of all SQL actions

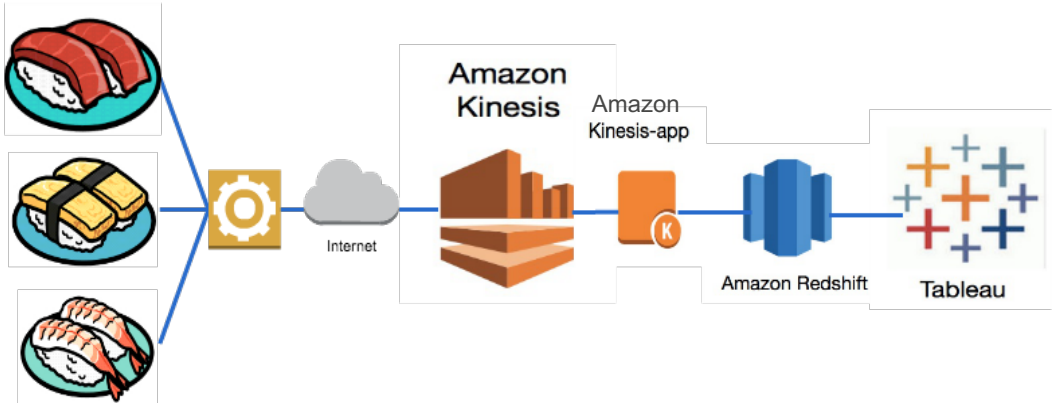
Audit of all configuration changes

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

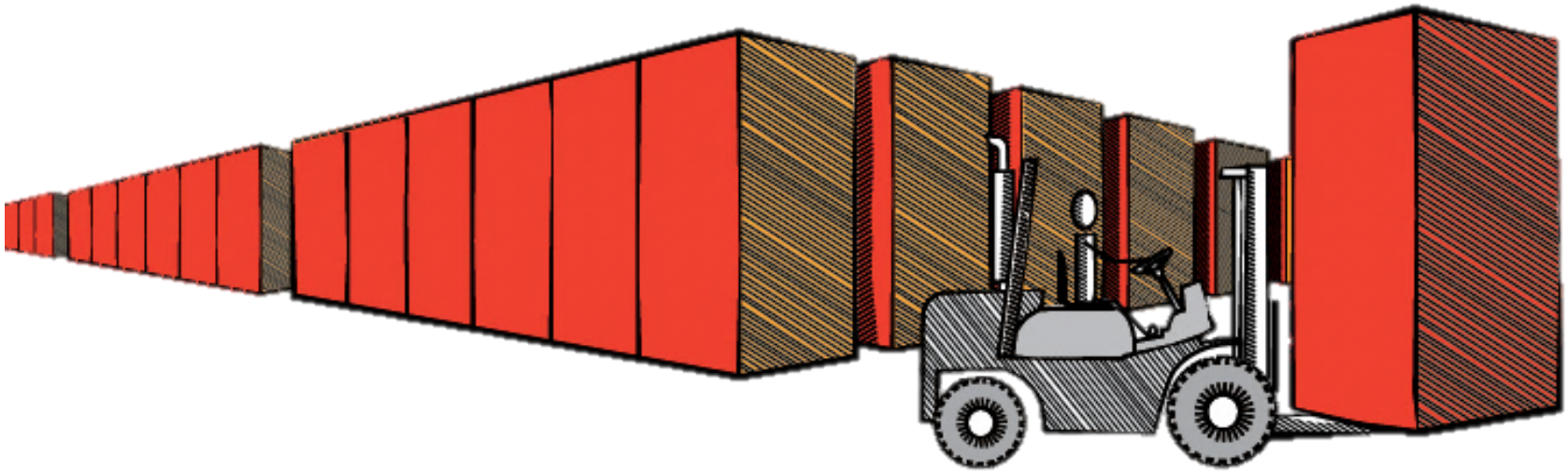
Big data does not mean batch

- 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

Q&A