AWS Big Data Platform

Agenda Overview

- Introduction to Big Data @ AWS
- Data Collection and Storage
- Real-time Event Processing
- Analytics (incl Machine Learning)
- Open Q&A Roundtable

Global Footprint

Everyday, AWS adds enough new server capacity to support Amazon.com when it was a $7 billion global enterprise.

Over 1 million active customers across 190 countries

800+ government agencies

3,000+ educational institutions

11 regions

28 availability zones

52 edge locations

IaaS Magic Quadrant

"AWS is the overwhelming market share leader, with more than 5X the compute capacity in use than the aggregate total of the other 14 providers."

Gartner

LearnOA® To The Next Level...

Enterprise Applications

- Virtual Desktop
- Sharing & Collaboration

Platform Services

Analytics

- Hadoop
- Real-time Streaming Data
- Data Warehouse
- Data Pipelines

App Services

- Queuing & Notifications
- Workflow
- App streaming
- Transcoding
- Email
- Search

Deployment & Management

- One-click web app deployment
- Dev/ops resource management
- Resource Templates

Mobile Services

- Identity
- Sync
- Mobile Analytics
- Push Notifications

Administration & Security

- Identity Management
- Access Control
- Usage Auditing
- Key Storage
- Monitoring And Logs

Core Services

- Compute (VMs, Auto-scaling and Load Balancing)
- Storage (Object, Block and Archival)
- CDN
- Databases (Relational, NoSQL, Caching)
- Networking (VPC, DX, DNS)

Infrastructure

- Regions
- Availability Zones
- Points of Presence

Broad & Deep Core Services

Compute

- Virtual Servers
- Containers
- Event-driven Compute Functions
- Auto Scaling
- Load Balancing

Storage & Content Delivery

- Object Storage
- Block Storage
- File System Storage
- Archive Storage
- CDN

Databases

- Relational
- NoSQL
- Caching

Networking

- Virtual Private Cloud
- Direct Connections
- DNS

Administration & Security

- Identity Management
- Access Control
- Usage & Resource Auditing
- Key Storage & Management
- Monitoring & Logs
- Service Catalog

Rich Platform Services

Analytics

- Hadoop
- Real-time
- Machine Learning
- Data Warehouse
- Data Pipelines

Application Services

- Queueing
- Workflow
- App Streaming
- Transcoding
- Email
- Search

Deployment & Management

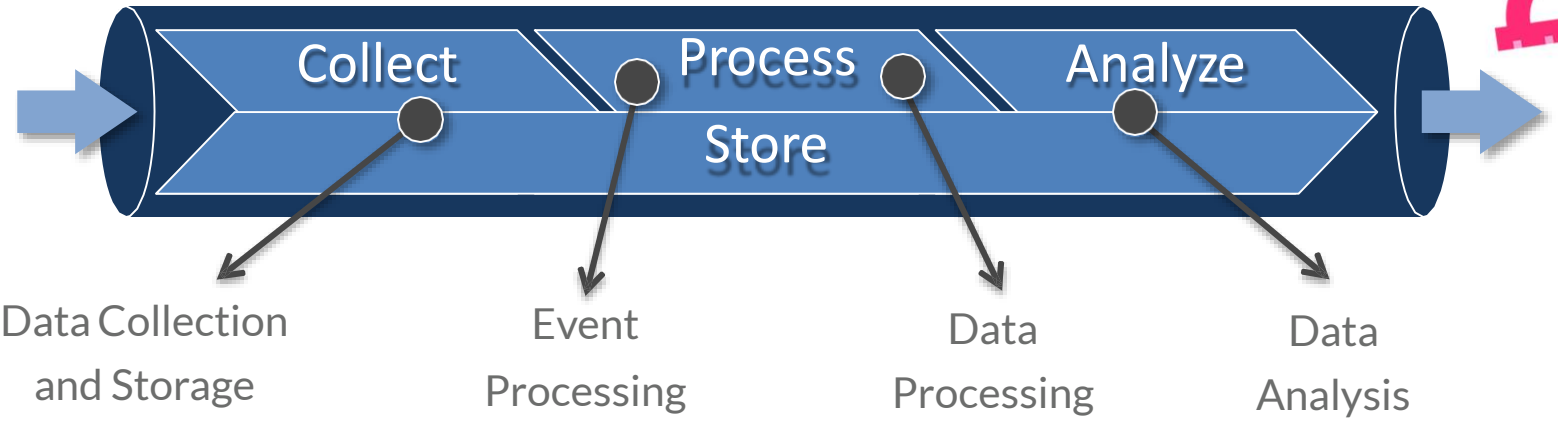
- 1-Click Web App Deployment
- Dev/Ops Resource Management
- Resource Templates
- Code Deployment
- Continuous Integration Tool
- Source Code Management

Mobile & Devices

- Identity
- Sync
- Mobile Analytics
- Notifications

Big Data Pipeline

Primitive Patterns



Primitive Patterns

Data Collection and Storage

- S3
- Kinesis
- DynamoDB
- RDS (Aurora)

Event Processing

- AWS Lambda
- KCL Apps

Data Processing

- EMR
- Redshift
- Machine Learning

Data Analysis

Primitive Patterns

Data Collection and Storage

Event Processing

Data Processing

Data Analysis

S3

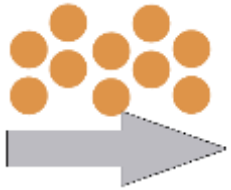
Kinesis

DynamoDB

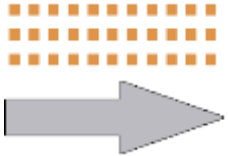
RDS (Aurora)

Data Collection and Storage

Apps Devices Logging Frameworks



File: media, log files (sets of records)



Stream: records (eg: device stats)

Transactional: database reads/writes

AWS services – data collection and storage

S3

Kinesis

DynamoDB RDS

(Aurora)

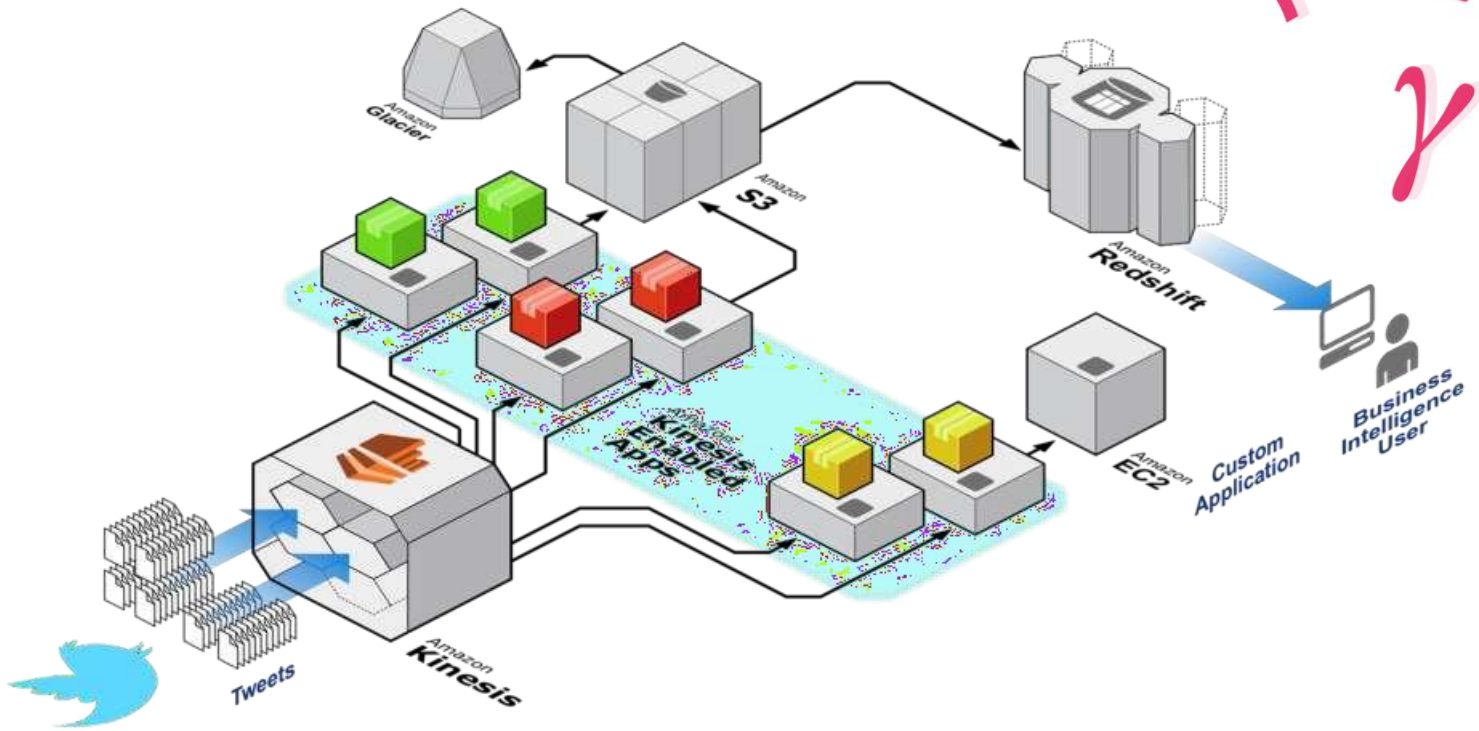
Benefits of Streamlined Data Collection

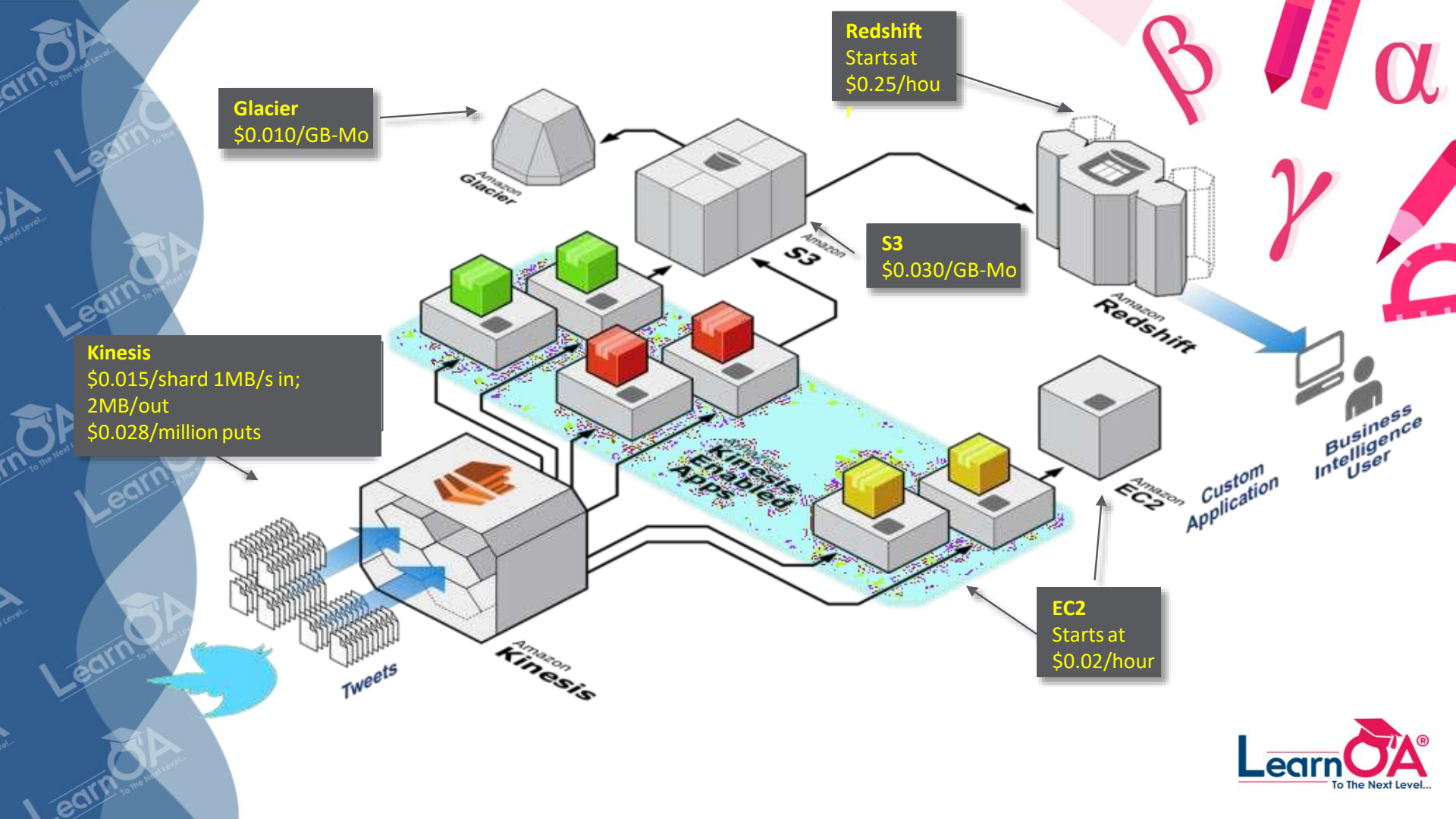
Increase velocity of data

- *Upgrade* existing applications to log records rather than files – driven by need for greater agility
- *Build* new applications that are designed for streaming data from the outset

  Example:

social media analytics (reference architecture)


Amazon Glacier
Amazon S3
Amazon Redshift
Business Intelligence User
Amazon Kinesis Enabled Apps
Amazon EC2
Custom Application
Tweets
Amazon Kinesis


Redshift
Starts at
$0.25/hou
Glacier
$0.010/GB-Mo
Amazon
Glacier
Amazon
S3
S3
$0.030/GB-Mo
Amazon
Redshift
Kinesis
$0.015/shard 1MB/s in;
2MB/out
$0.028/million puts
Amazon
Kinesis
Enabled
Apps
Business
Intelligence
User
Amazon
EC2
Custom
Application
Tweets
Amazon
Kinesis
EC2
Starts at
$0.02/hour

Cost & Scale

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

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

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

Kinesis cost is $0.765/hour

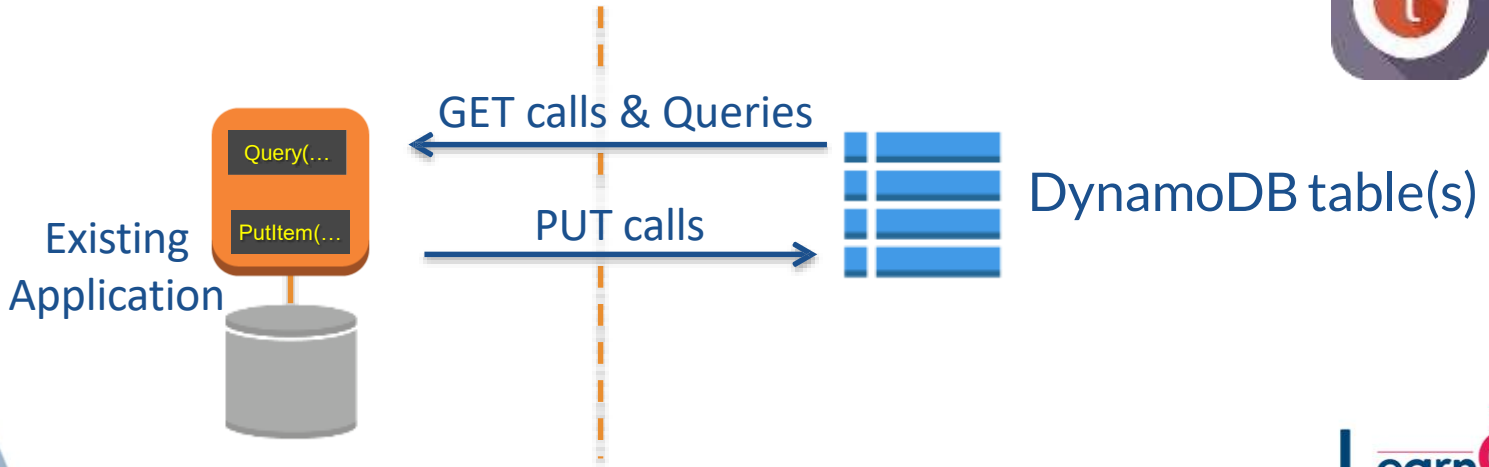
Redshift cost is $0.850/hour (for a 2TB node)

S3 cost is $1.28/hour (no compression)

Total: $2.895/hour

Benefits of Streamlined Data Collection

- Instrument existing applications
- Inject code to log activity – “new big data”
- Example: WAPO Labs Social Reader (now Trove)



Benefits of Streamlined Data Collection

Increase data granularity

Customers

Devices

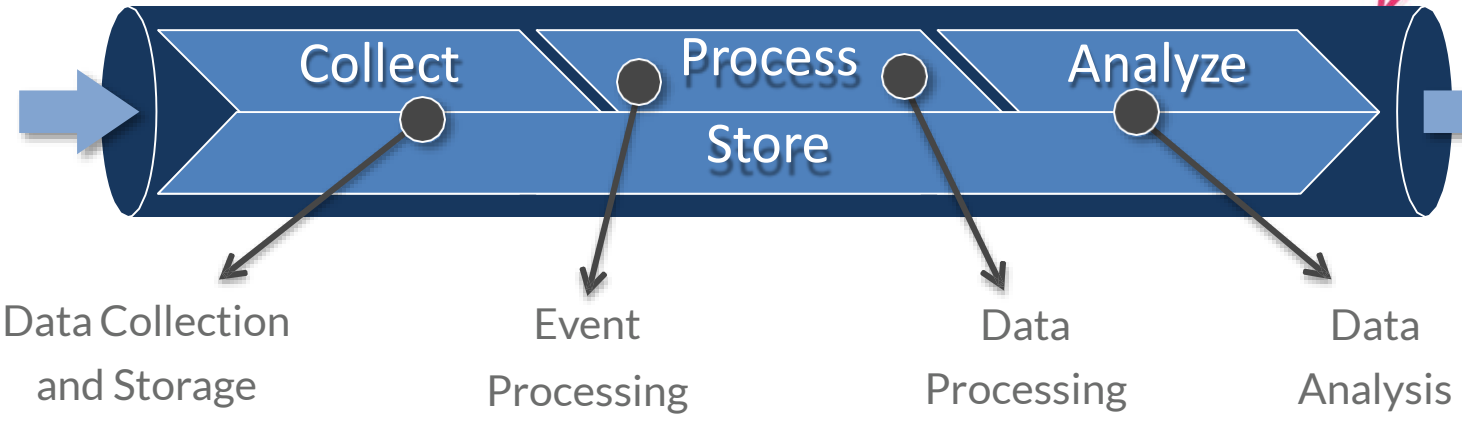
Data Items

Item Size

Frequency

Challenge: compounding scale Benefit: improved data quality

Primitive Patterns



AWS Lambda

KCL Apps

Event Processing – Enabling Capabilities

S3 Event Notifications

Kinesis stream

DynamoDB Streams

KCL Apps

AWS Lambda

Real-Time Event Processing

- Event-driven programming
- Trigger activities based on real-time input

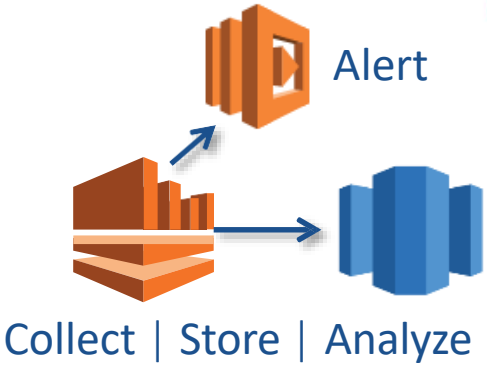
Examples:

- Proactively detect hardware errors in device logs
- Identify fraud from activity logs
- Monitor performance SLAs
- Notify when inventory drops below a threshold

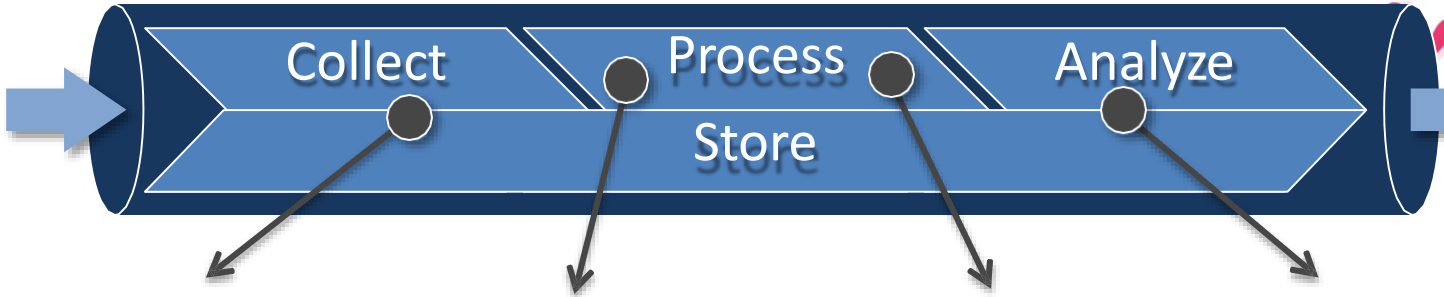
Benefits of Event Processing

- Build / add real-time events
  - Take action between data collection and analytics
    - ✓ Alerts and notifications, performance and security
    - ✓ Automated data enrichment (eg: aggregations)

- De-couple application modules
  - Streamline development and maintenance
  - Increase agility
    - ✓ MVP + iterate on discrete components



Primitive Patterns



Data Collection and Storage

Event Processing

Data Processing

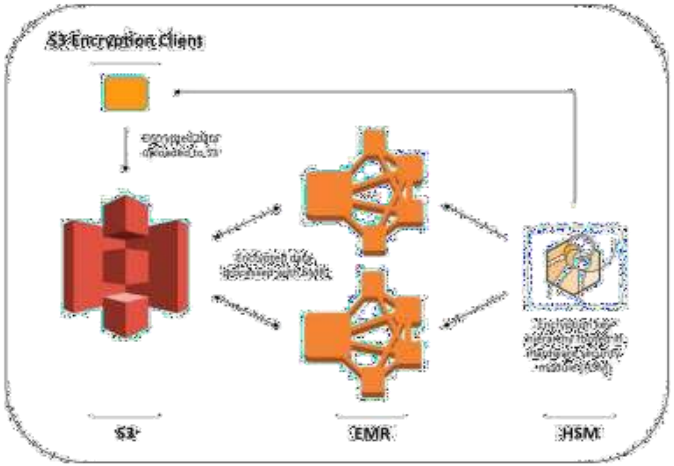
Data Analysis

EMR

Redshift

Machine Learning

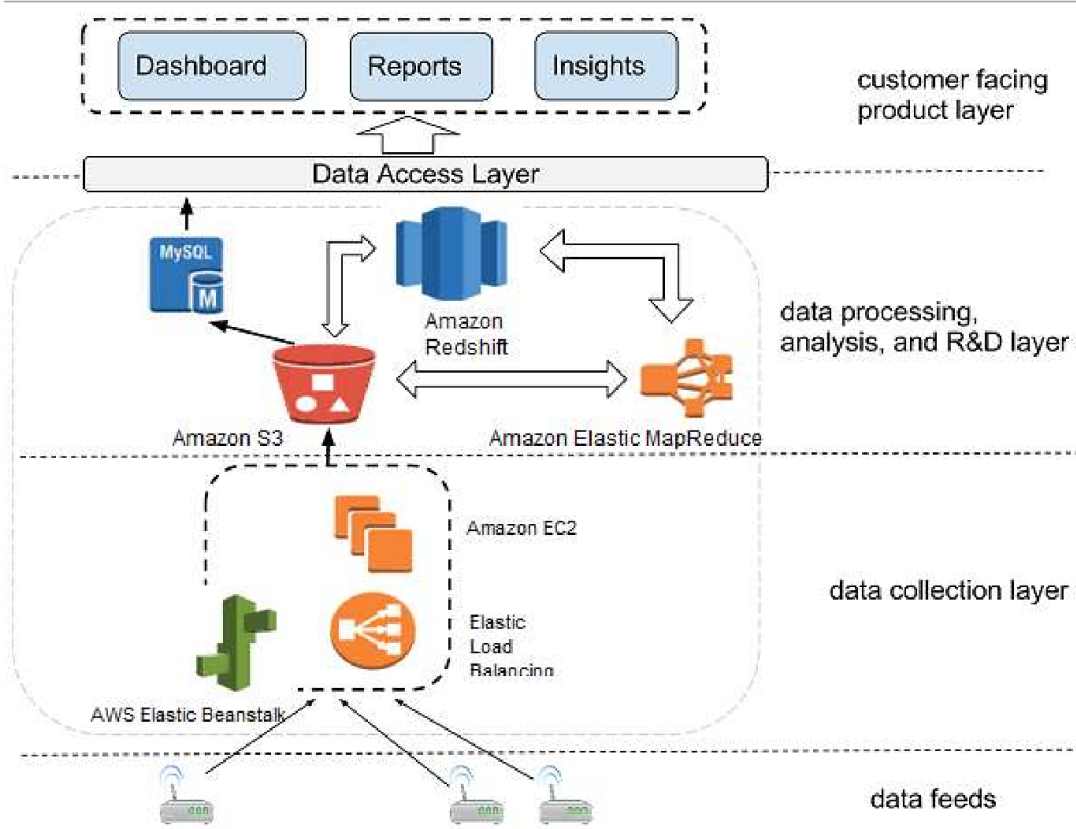
NASDAQ



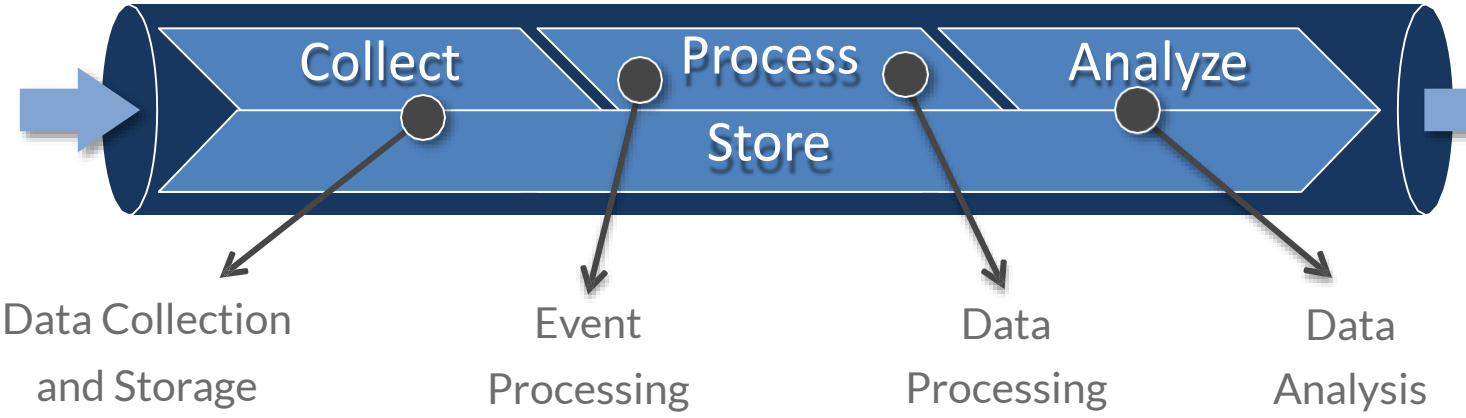
- 5.5B Records are loaded to Amazon Redshift every day
- Security Requirements for Client Side Encryption
- Historical Data - HDFS became too expensive
  - S3 + EMR to the Rescue

euclid

- Retail and POS Analytics
- Process 10's of TB in hours vs. 2 weeks
- 80-90% reduction in costs



Big Data Use Cases



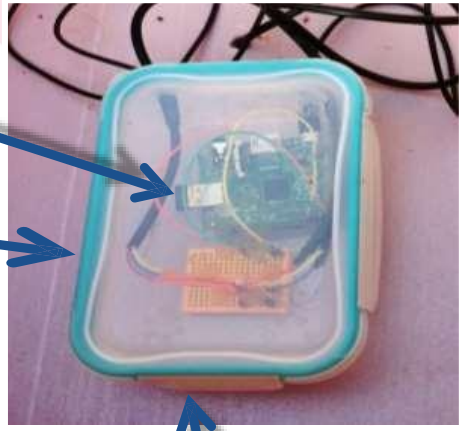
Internet of Things
Digital Advertising
Online Gaming

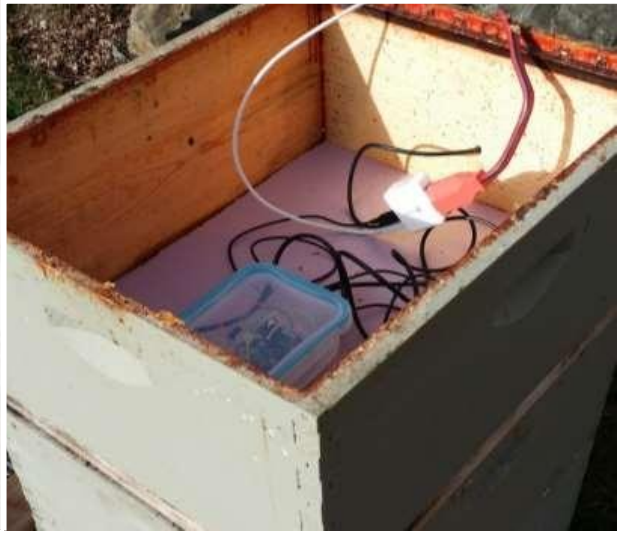
Log Analytics
Customer Value Scoring
Personalization Engine

TempTracker
bee hive monitoring in the AWS cloud

network card

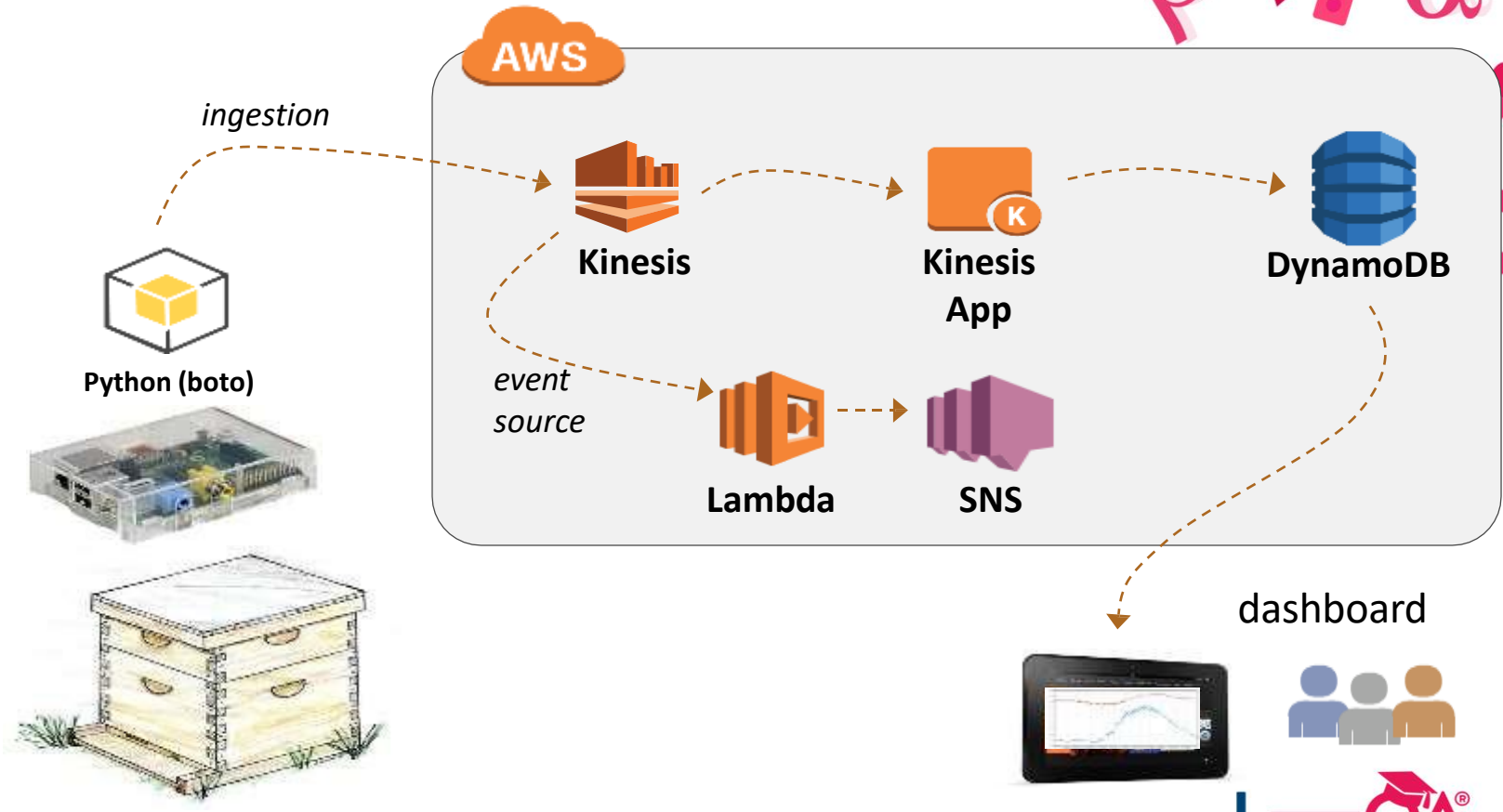
waterproof housing





temperature sensors

TempTracker: IoT sensor ingestion example

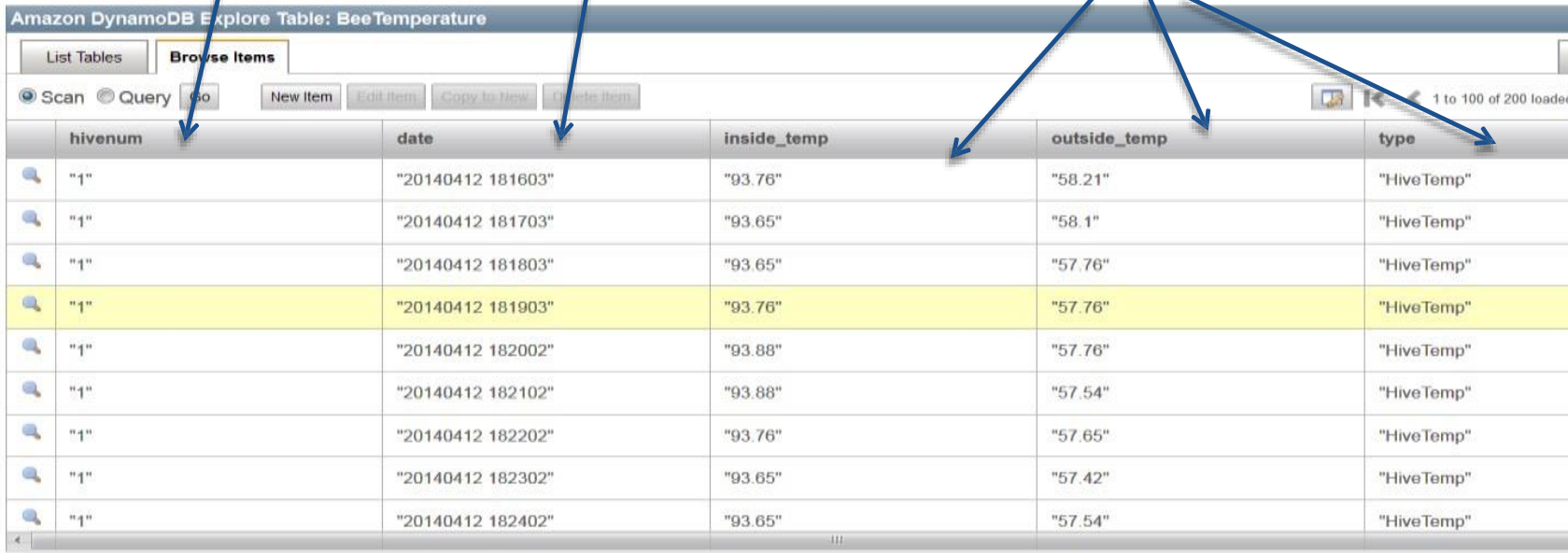


DynamoDB schema

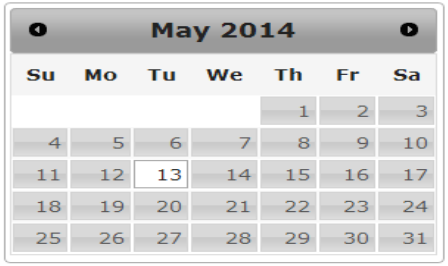
hash

range

attributes



hivenum	date	inside_temp	outside_temp	type
"1"	"20140412 181603"	"93.76"	"58.21"	"HiveTemp"
"1"	"20140412 181703"	"93.65"	"58.1"	"HiveTemp"
"1"	"20140412 181803"	"93.65"	"57.76"	"HiveTemp"
"1"	"20140412 181903"	"93.76"	"57.76"	"HiveTemp"
"1"	"20140412 182002"	"93.88"	"57.76"	"HiveTemp"
"1"	"20140412 182102"	"93.88"	"57.54"	"HiveTemp"
"1"	"20140412 182202"	"93.76"	"57.65"	"HiveTemp"
"1"	"20140412 182302"	"93.65"	"57.42"	"HiveTemp"
"1"	"20140412 182402"	"93.65"	"57.54"	"HiveTemp"

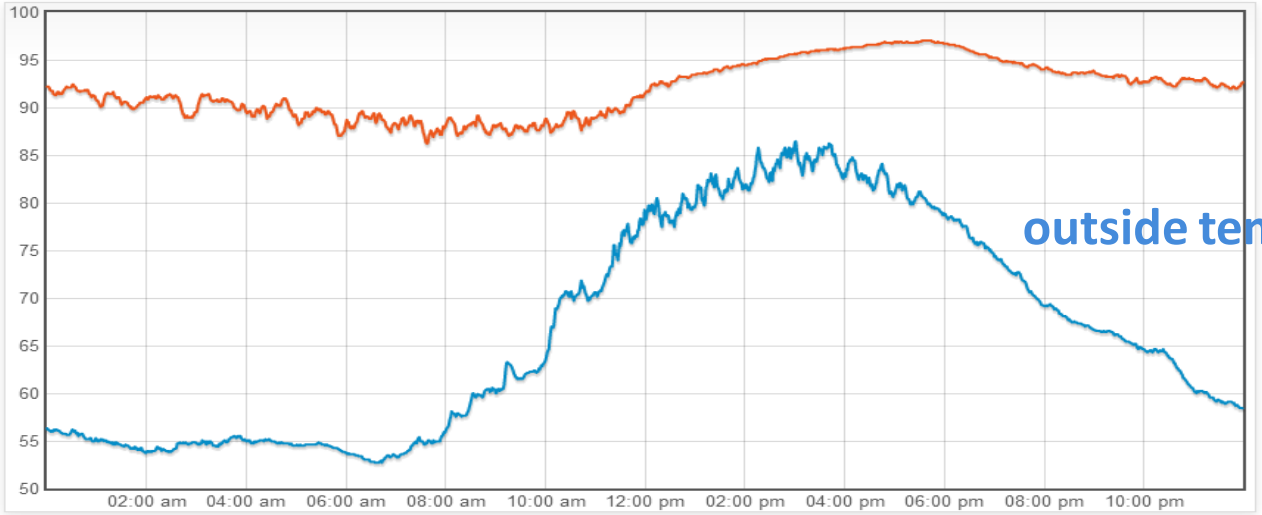


Welcome to Hive Temp Tracker

Honeybees have a remarkable ability to maintain temperature within a beehive. This is especially important throughout the baby bee rearing months. Special bees withing the hive-- known as heater bees-- have body temperatures are considerably higher than other bees in the colony. They use this heat to not only keep the hive warm but also control the social make-up within a colony.

Go get it

internal temperature



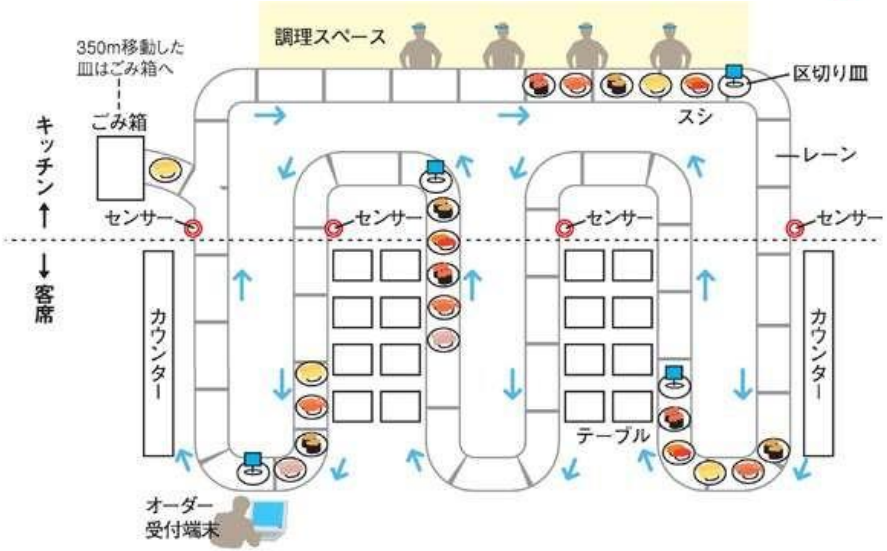
outside temperature

Big Data Case Study: Kaiten Sushiro

Kaiten Sushiro

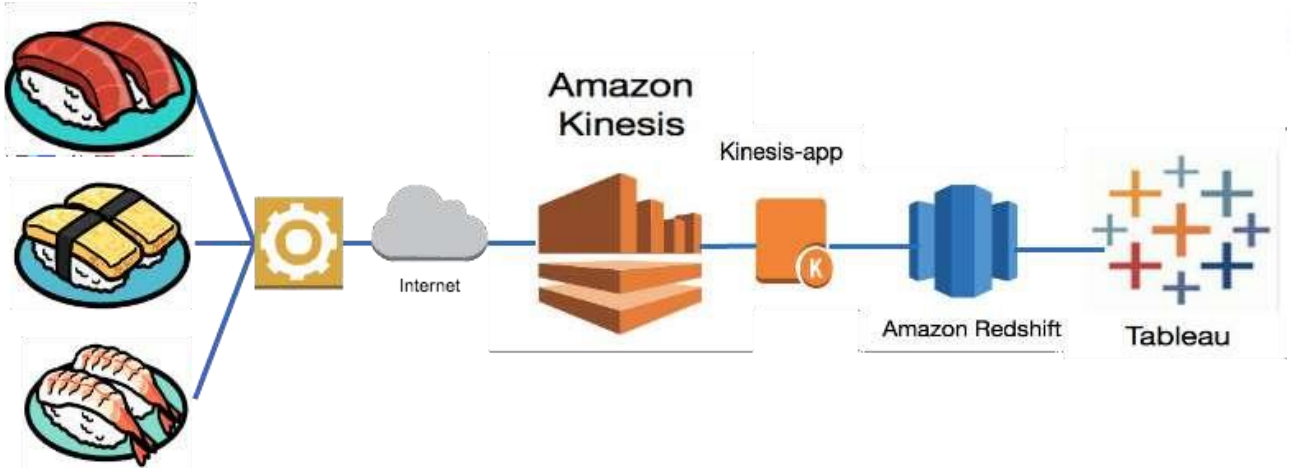
スシロー SUSHIRO

- Kaiten Sushi Chain restaurant
- Gathering sensor data into Kinesis



LearnQA To The Next Level...

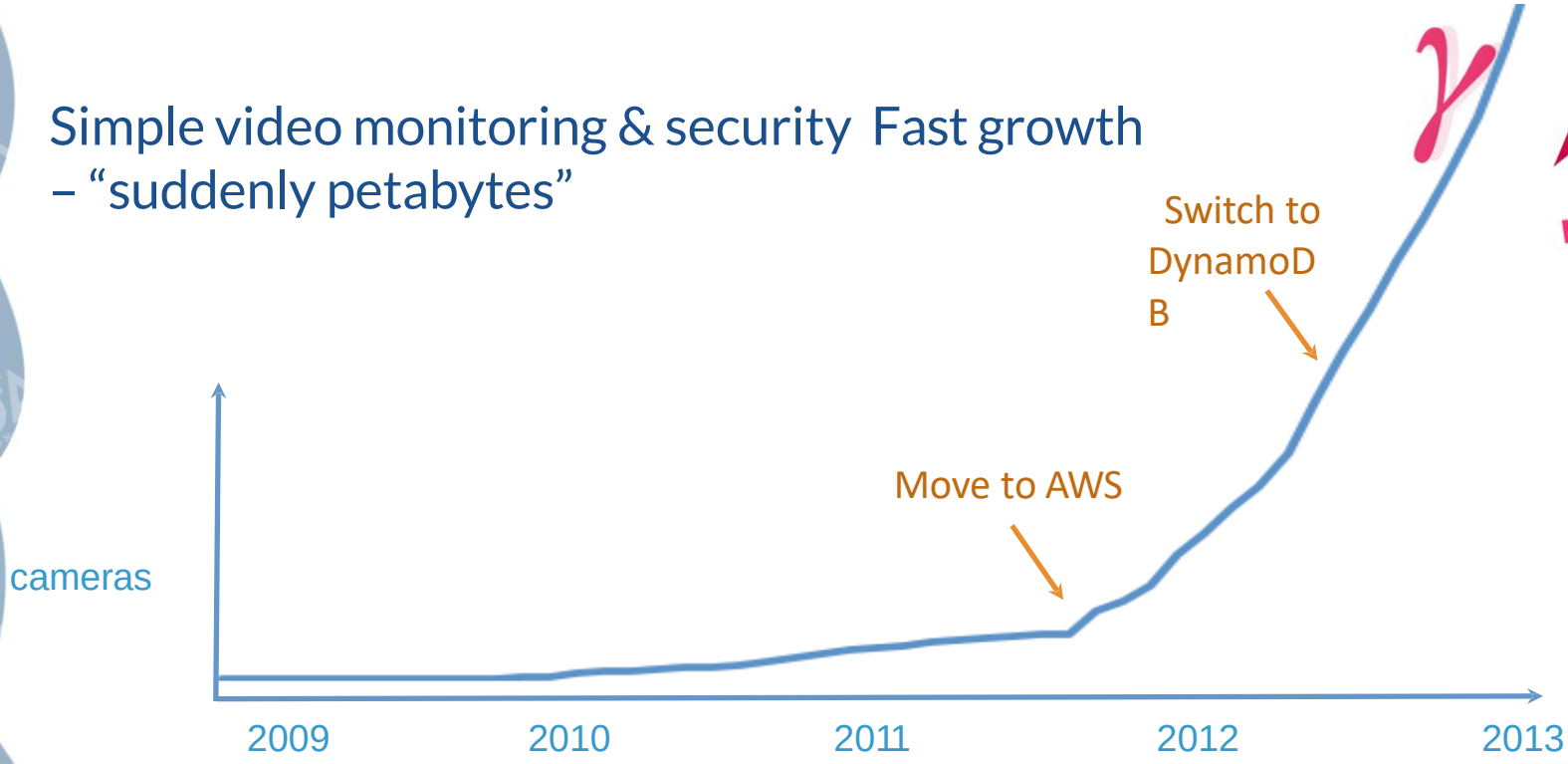
Kaiten Sushiro Data Flow



IoT / connected devices

dropcam

Simple video monitoring & security Fast growth – “suddenly petabytes”



LearnOA® To The Next Level...

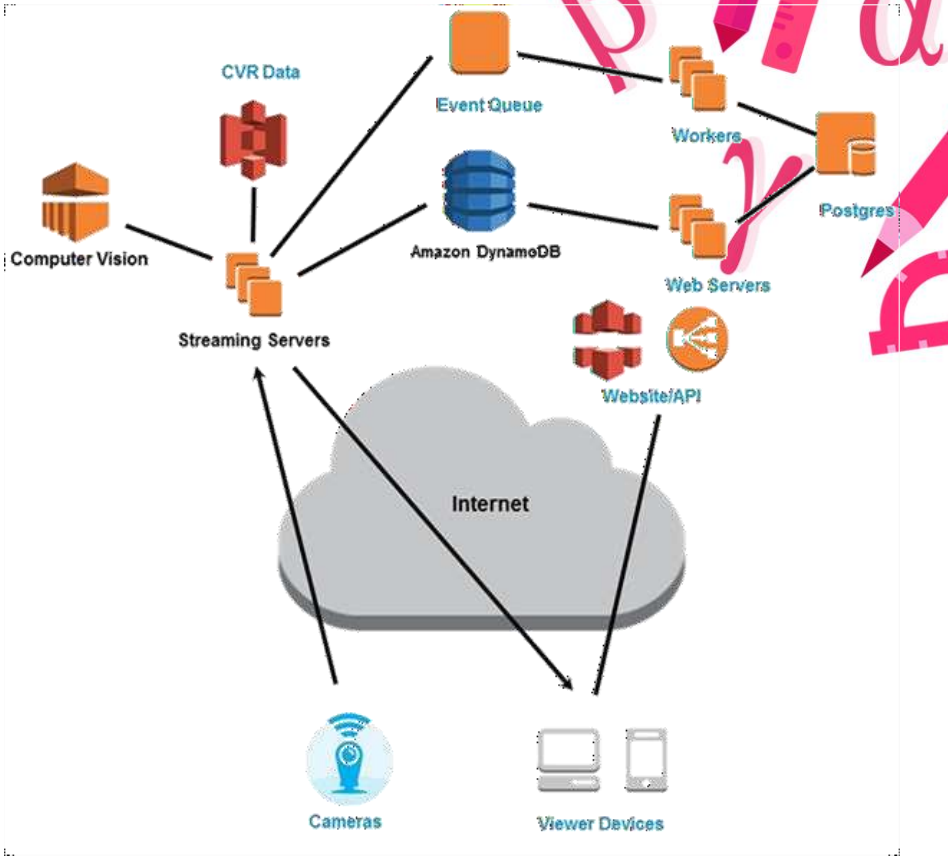
dropcam

EC2 (live streaming) S3 (CVR data)

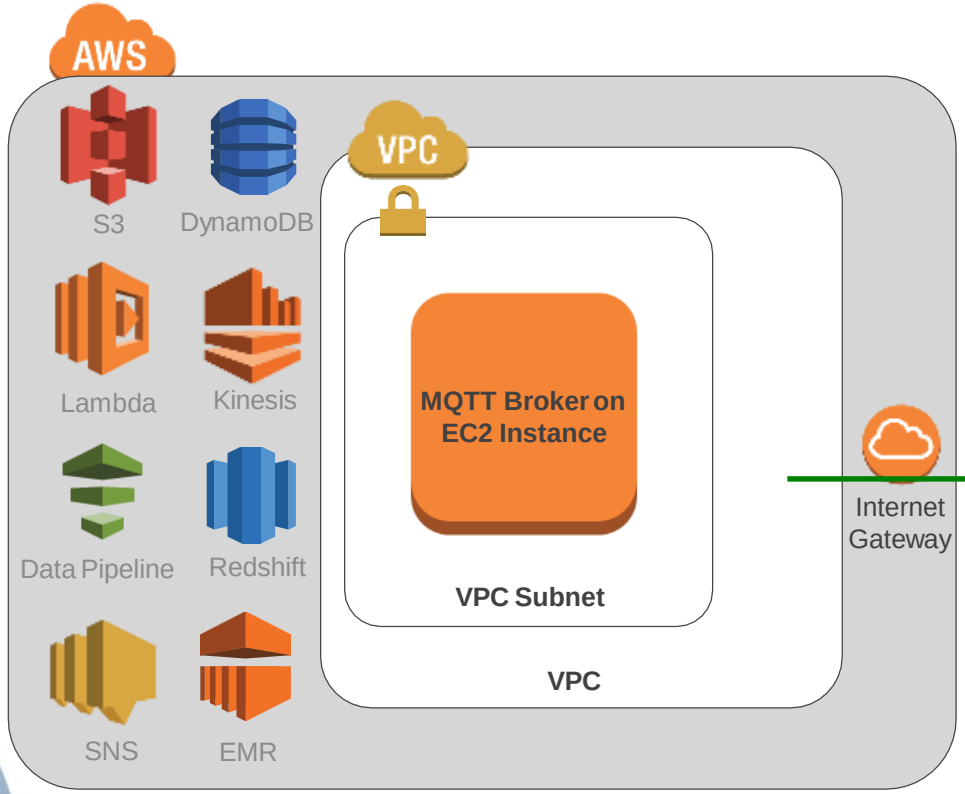
DynamoDB (meta data)

CloudFront (CDN)

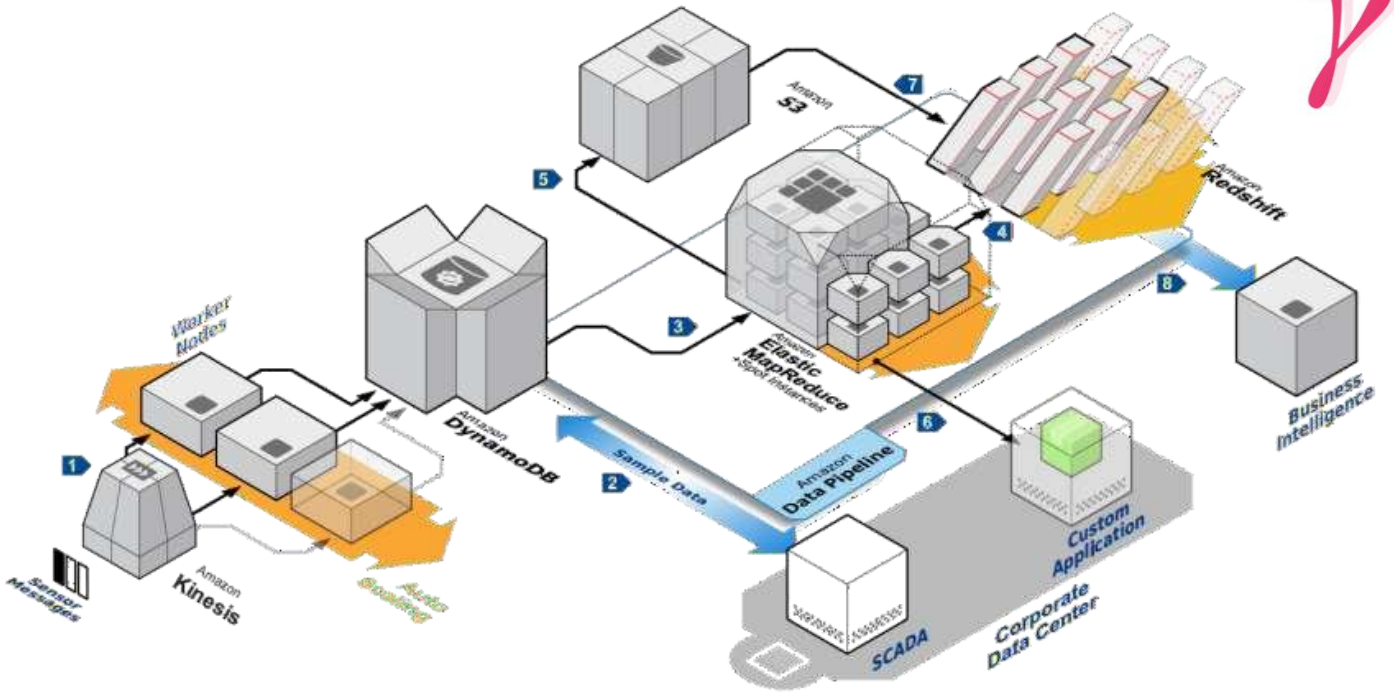
EMR (activity recognition)



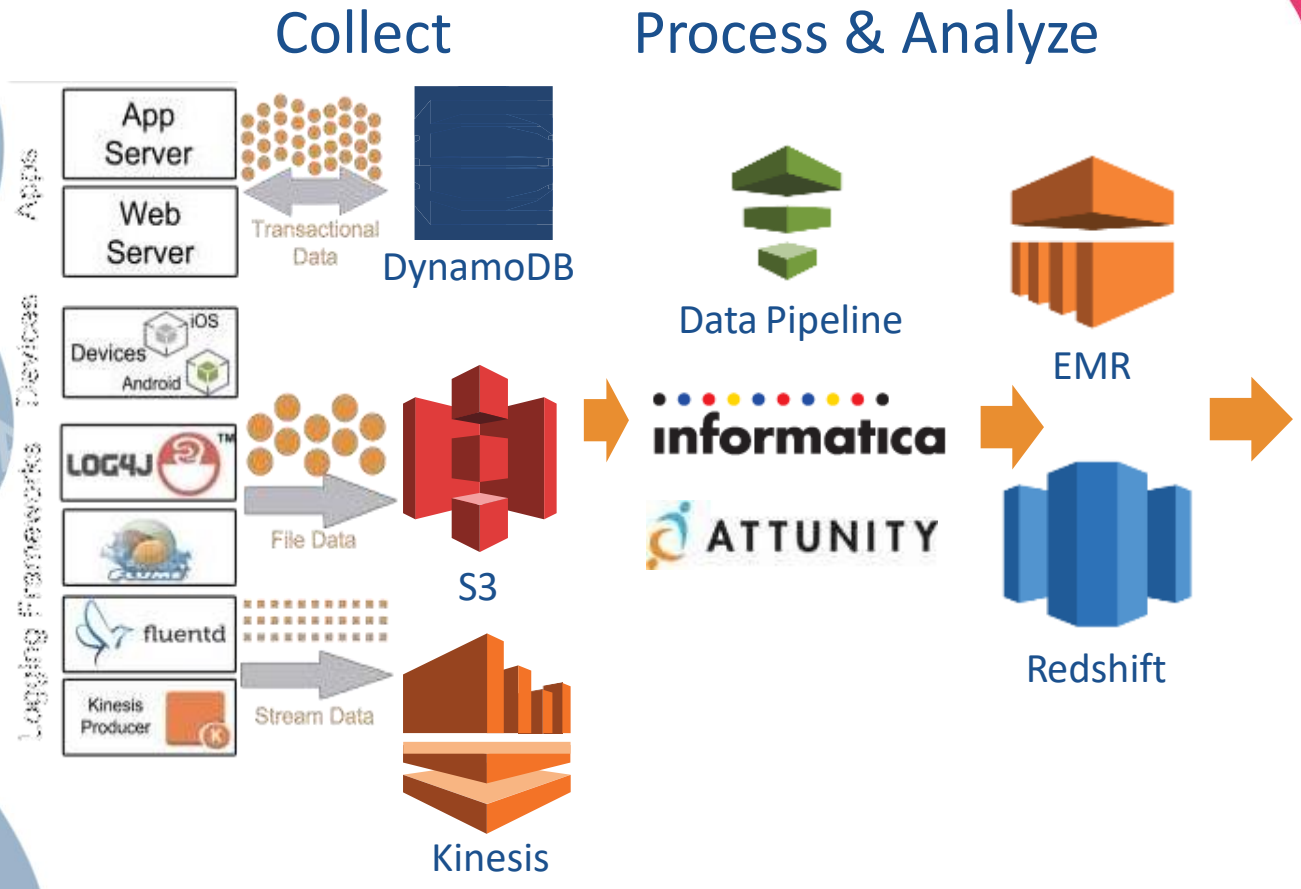
Applying Analytics to Connected Device Data



Backend Analytics Architecture for Connected Device Data



AWS big data ecosystem



AWS Professional Services Partnering in Your Journey

Technical Specialists

Specialty practices for AWS skills transfer, security, infrastructure architecture, application optimization, analytics, big data, and operational integration

Advisory Services

Portfolio strategy and planning, cost/benefit modeling, governance, change management and risk management as it relates to implementing the AWS platform

Collaboration

Working together with you and APN Premier Partners you already trust to provide you with access to all resources needed to realize breakthrough results

Proven Process

Best practices and patterns to help your teams get the foundation right, deploy and migrate workloads, and create a modern IT operating model to support your business

Big Data Partner Solutions

Solutions vetted by the AWS Partner Competency Program

Data Enablement	Data Analysis & Visualization	Infrastructure Intelligence	Advanced Analytics
Move, synchronize, cleanse, and manage data	Turn data into actionable insight, enhance decision making	Harness data generated from your systems and infrastructure	Anticipate future events and behaviors, conduct what-if analysis
ATTUNITY	looker	splunk>	SAP
informatica	MicroStrategy	sumologic	MAPR
snaplogic	TIBCO Jaspersoft		CIVIS ANALYTICS
	tableau		

big data service offers
Service expertise vetted by the AWS Partner Competency Program

AWS marketplace
Enterprise software store for business users who need simplified procurement

2,000+ product listings to browse, test and buy software

1-click deployment to launch, on multiple regions around the world

Pay-as-you-go pricing with no long term contracts required

awsmarketplace

Business Intelligence

TIBCO Spotfire, Logi Analytics, SAP Business Objects, MicroStrategy, looker, pentaho

Advanced Analytics

PredictionIO, RECICTEK, Atigeo, bigml, Digital Reasoning, ZEMENTIS

Database and Data Enablement

syncsort, informatica, ATTUNITY CloudBeam, SAP HANA, Couchbase, MarkLogic

LearnOA® To The Next Level...

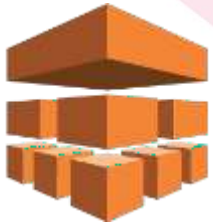
Amazon Machine Learning

Amazon Aurora

Smart Applications

- **e-commerce:** recommendations made based on your past purchases
- **finance:** alerts from your bank when they suspect fraudulent transactions
- **retail:** emails when items related to things you typically buy are on sale

Amazon Machine Learning



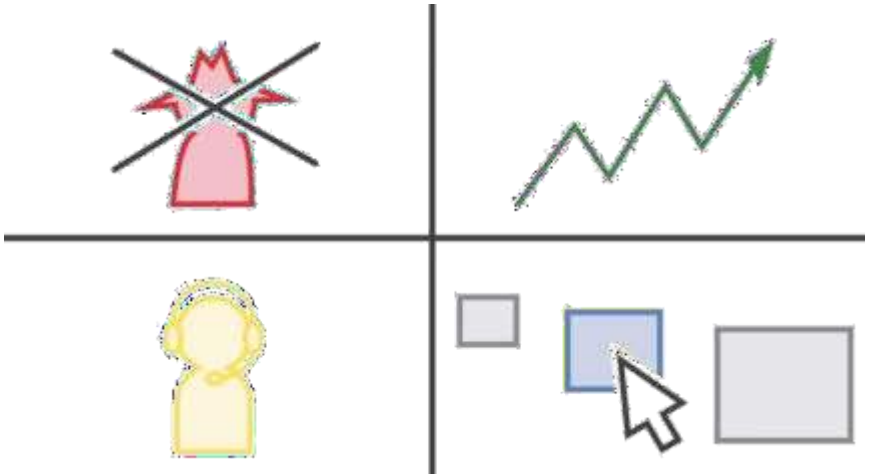
1. Build & Train Model
   - Create a datasource object (connect to Redshift, RDS, S3)
   - Explore and understand your data
   - Transform and train your model

2. Evaluate the Model & Optimize
   - Assess model quality
   - Fine-tune the model

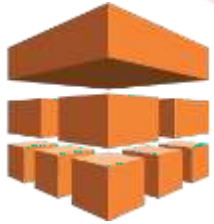
3. Retrieve Predictions
   - Batch: asynchronous, large volume prediction
   - Real-time: synchronous, single-item prediction

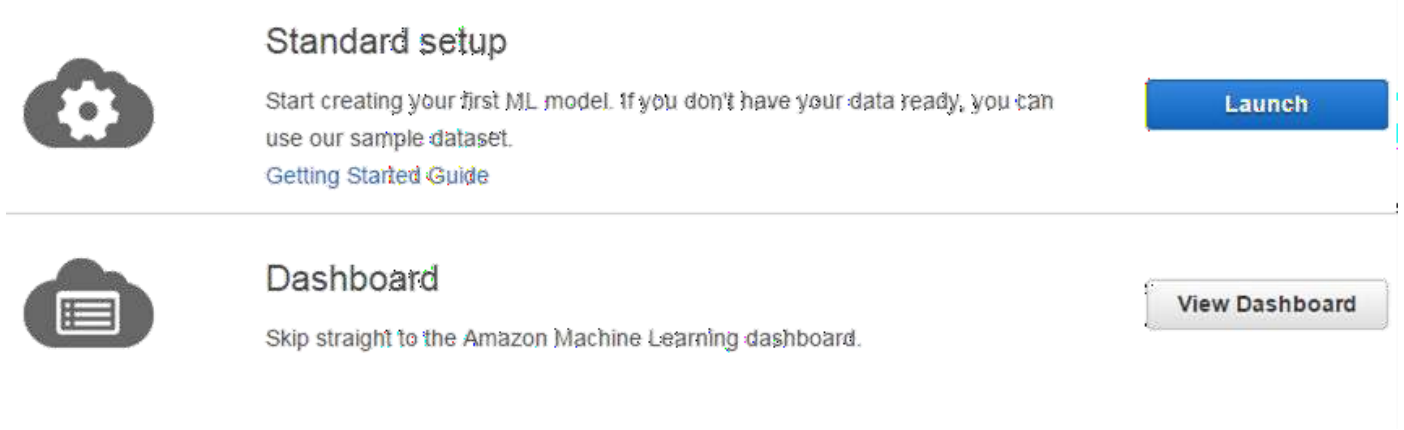
Amazon Machine Learning example use cases

- Fraud detection
- Demand forecasting
- Predictive customer support
- Click prediction
- Content personalization
- Document classification



Amazon Machine Learning Currently Available in US-East-1



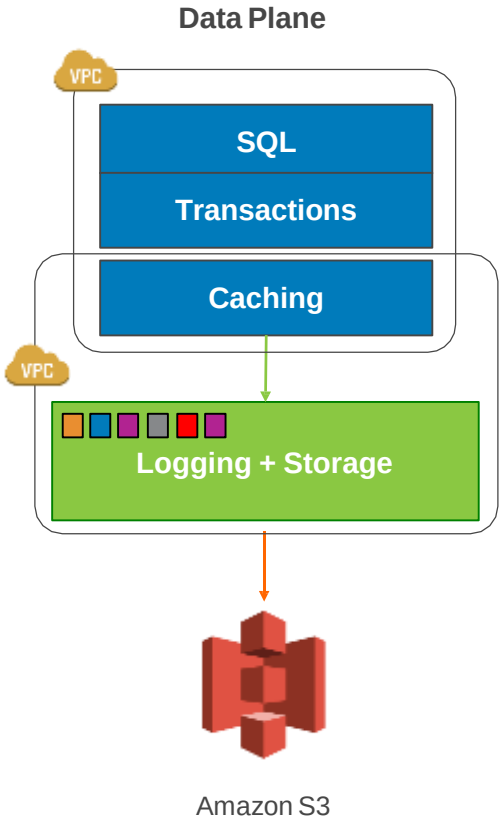


Amazon Aurora

Amazon's New Relational Database Engine

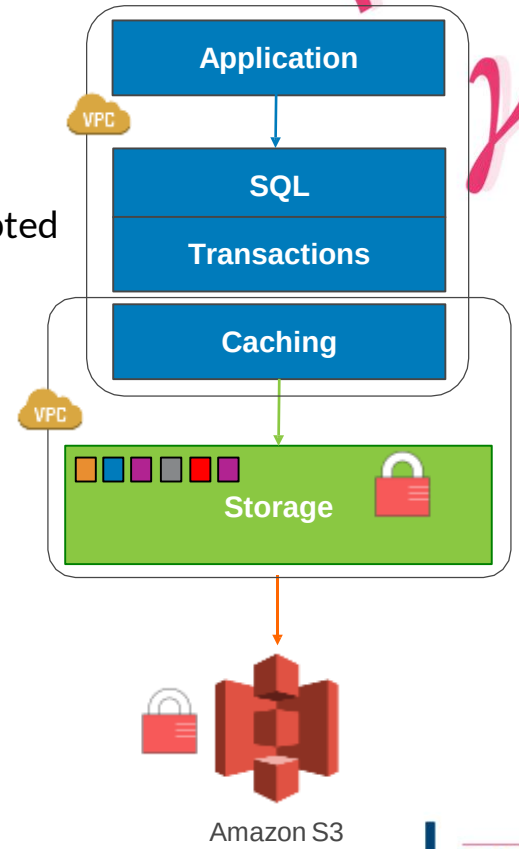
A Service-Oriented Architecture Applied to the database

- Moved the logging and storage layer into a multi-tenant, scale-out database-optimized storage service
- Integrated with other AWS services like Amazon EC2, Amazon VPC, Amazon DynamoDB, Amazon SWF, and Amazon Route 53 for control plane operations
- Integrated with Amazon S3 for continuous backup with 99.999999999% durability



Simplify Data Security

- Encryption to secure data at rest
  - AES-256; hardware accelerated
  - All blocks on disk and in Amazon S3 are encrypted
  - Key management via AWS KMS
- SSL to secure data in transit
- Network isolation via Amazon VPC by default
- No direct access to nodes
- Supports industry standard security and data protection certifications

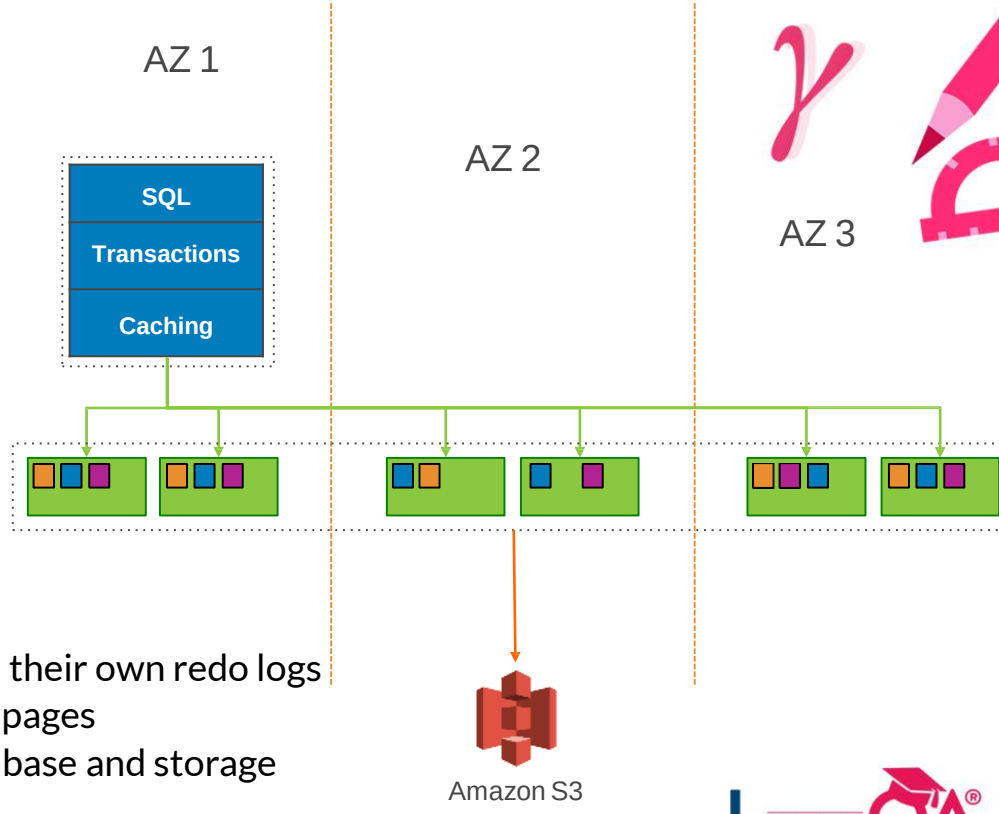


Simplify Storage Management

- Read replicas are available as failover targets—no data loss
- Instantly create user snapshots—no performance impact
- Continuous, incremental backups to S3
- Automatic storage scaling up to 64 TB—no performance or availability impact
- Automatic restriping, mirror repair, hot spot management, encryption

Aurora Storage

- Highly available by default
- 6-way replication across 3 AZs
- 4 of 6 write quorum
  - Automatic fallback to 3 of 4 if an AZ is unavailable
- 3 of 6 read quorum



SSD, scale-out, multi-tenant storage

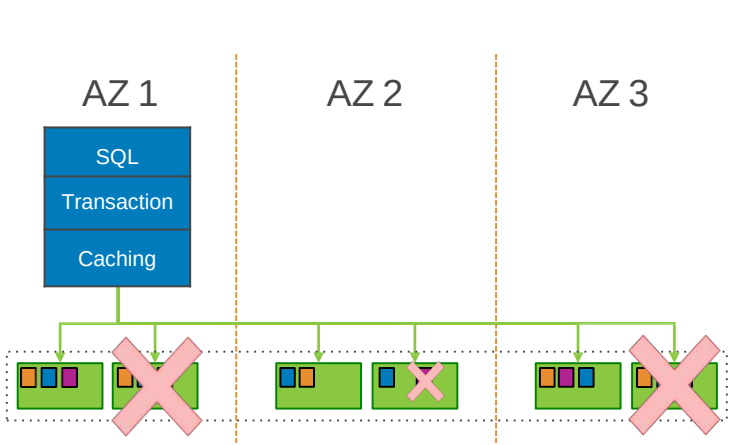
- Seamless storage scalability
- Up to 64 TB database size
- Only pay for what you use

Log-structured storage

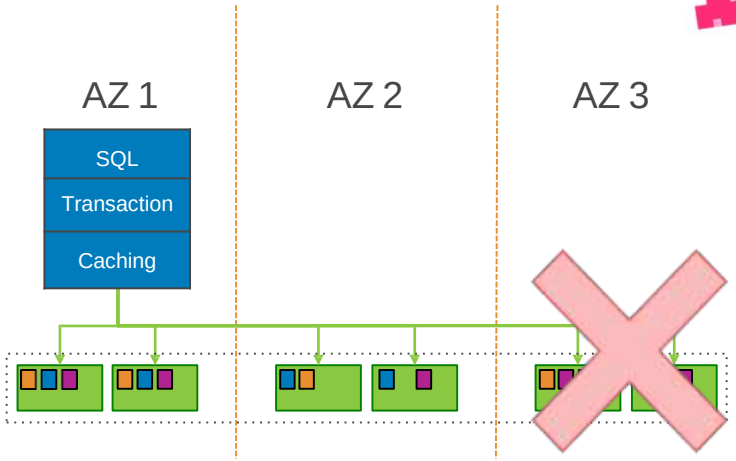
- Many small segments, each with their own redo logs
- Log pages used to generate data pages
- Eliminates chatter between database and storage

Self-healing, fault-tolerant

- Lose two copies or an AZ failure without read or write availability impact
- Lose three copies without read availability impact
- Automatic detection, replication, and repair



Read availability



Read and write availability

Instant crash recovery

Traditional databases

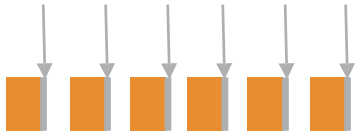
- Have to replay logs since the last checkpoint
- Single-threaded in MySQL; requires a large number of disk accesses

Crash at T_0 requires a re-application of the SQL in the redo log since last checkpoint

Amazon Aurora

- Underlying storage replays redo records on demand as part of a disk read
- Parallel, distributed, asynchronous

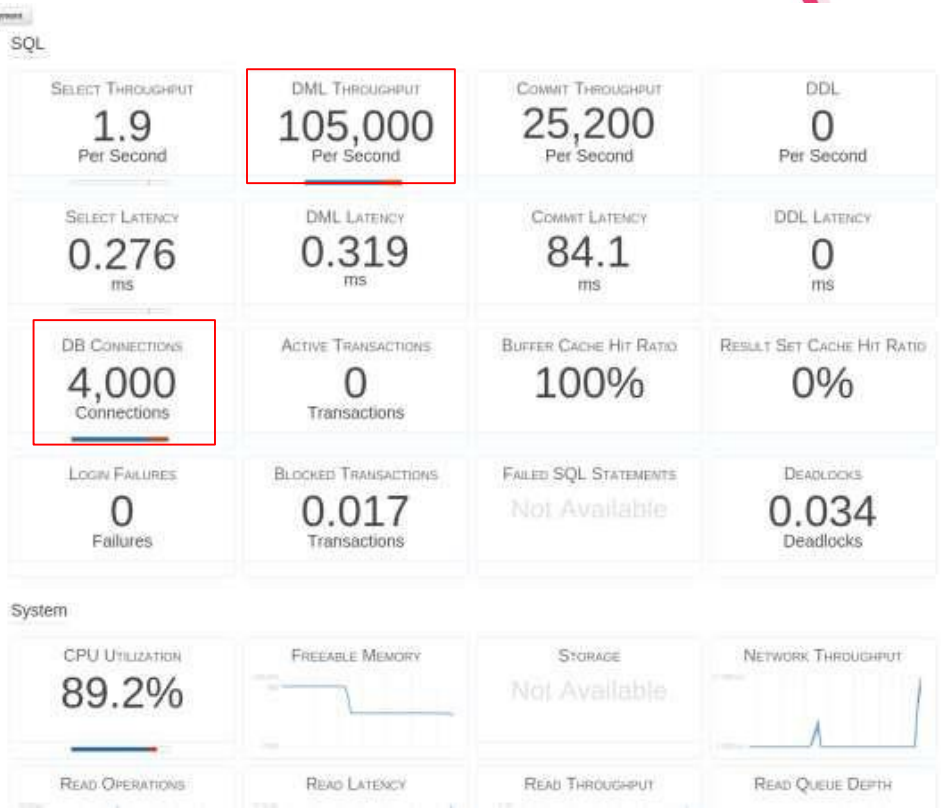
Crash at T_0 will result in redo logs being applied to each segment on demand, in parallel, asynchronously



T_0

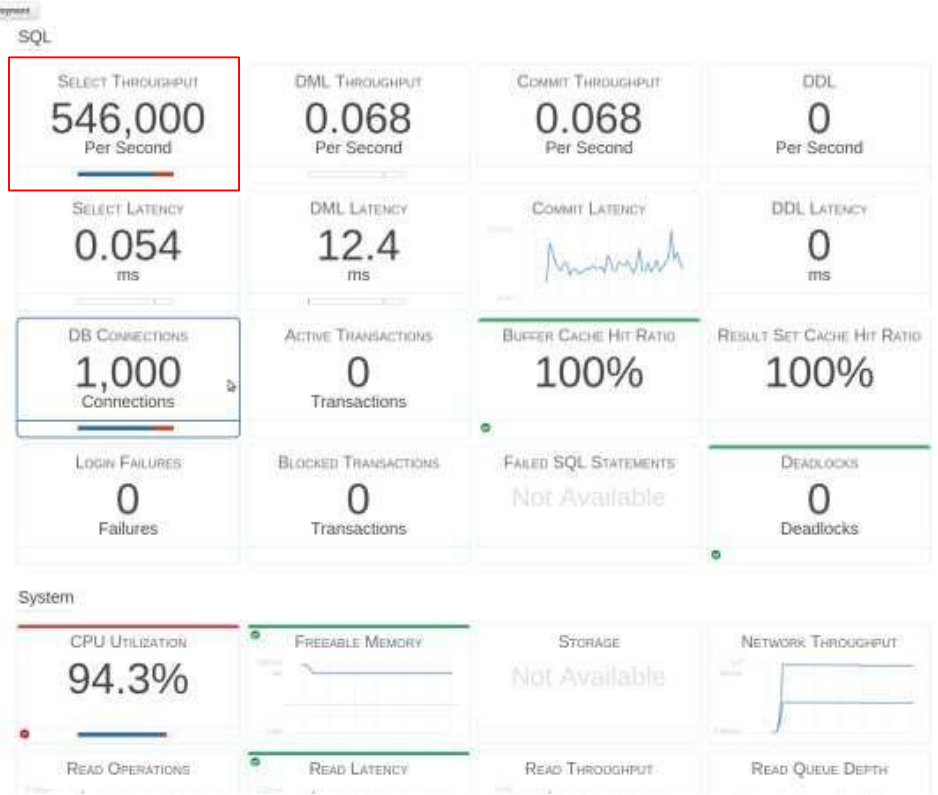
Write performance (console screenshot)

- MySQL Sysbench
- R3.8XL with 32 cores and 244 GB RAM
- 4 client machines with 1,000 threads each



Read performance (console screenshot)

- MySQL Sysbench
- R3.8XL with 32 cores and 244 GB RAM
- Single client with 1,000 threads



Read replica lag (console screenshot)

- Aurora Replica with 7.27 ms replica lag at 13.8 K updates/second
- MySQL 5.6 on the same hardware has ~2 s lag at 2 K updates/second

Aurora – current state

- Sign up for preview access at: https://aws.amazon.com/rds/aurora/preview
- Now available in US West (Oregon) and EU (Ireland), in addition to US East (N. Virginia)
- Thousands of customers already in the limited preview
- Unlimited preview: accepting all requests from late May
- Full service launch in the coming months

AWS big data platform

- **Choice** – platform breadth supports many use cases
- **Specialization** – optimal application experiences
- **Managed Services** – eliminate undifferentiated effort


S3


Kinesis


DynamoDB


RDS (Aurora)


AWS Lambda


KCL Apps


EMR


Redshift


Machine Learning

Thank You!