AWS re:Invent

Introducing Amazon Kinesis

Managed Service for Real-time Big Data Processing

Ryan Waite, GM Data Services

Adi Krishnan, Product Manager

November 13, 2013

amazon web services

Introducing Amazon Kinesis

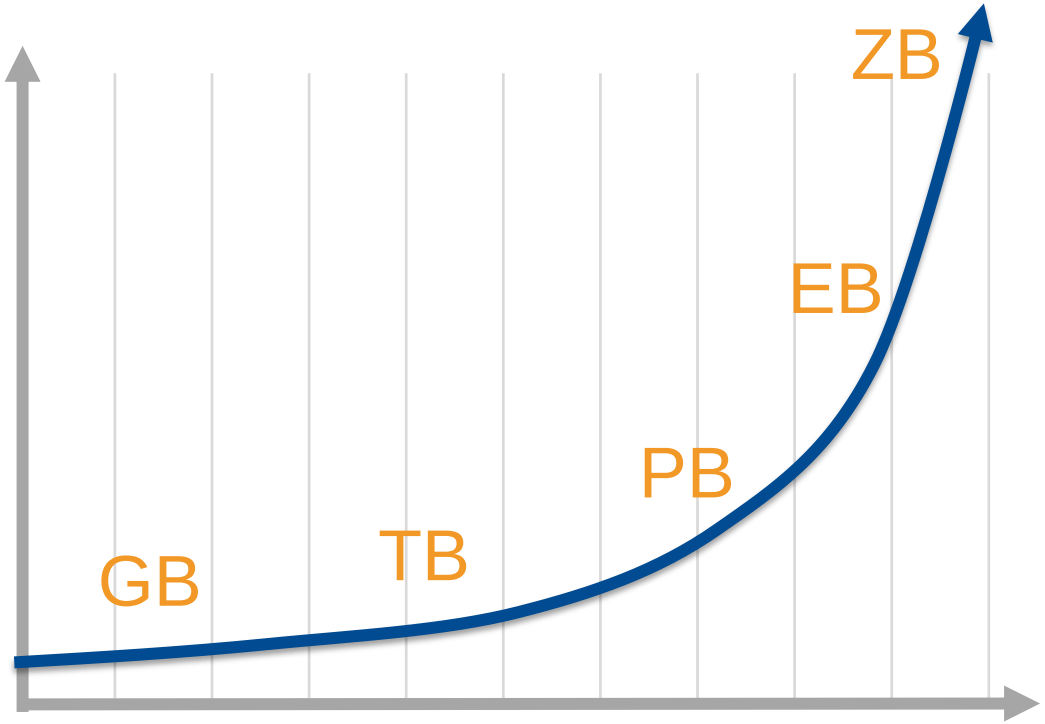
Managed service for real-time processing of big data

- Moving from Batch to Continuous, Real-time Processing
- How Does Real-time Processing Fit in with Other Big Data Solutions?
- Amazon Kinesis Features & Benefits
- Amazon Kinesis Key Concepts
- Customer Use Cases & Patterns

Why Real-Time Processing?

Unconstrained Data Growth

Big Data is now moving fast …



- **IT/ Application server logs**
  IT Infrastructure logs, Metering, Audit logs, Change logs
- **Web sites / Mobile Apps/ Ads**
  Clickstream, User Engagement
- **Sensor data**
  Weather, Smart Grids, Wearables
- **Social Media, User Content**
  450MM+ Tweets/day

No Shortage of Big Data Processing Solutions

Right Toolset for the Right Job

- Common Big Data Processing Approaches
  - Query Engine Approach (Data Warehouse, YesSQL, NoSQL databases)
    - Repeated queries over the same well-structured data
    - Pre-computations like indices and dimensional views improve query performance
  - Batch Engines (Map-Reduce)
    - Semi-structured data is processed once or twice
    - The “query” is run on the data. There are no pre-computations.
- Streaming Big Data Processing Approach
  - Real-time response to content in semi-structured data streams
  - Relatively simple computations on data (aggregates, filters, sliding window, etc.)
  - Enables data lifecycle by moving data to different stores / open source systems

Big Data : Served Fresh

Internal AWS experiences provided inspiration

Big Data

- Hourly server logs: how your systems were misbehaving an hour ago
- Weekly / Monthly Bill: What you spent this past billing cycle?
- Daily customer-preferences report from your website's click stream: tells you what deal or ad to try next time
- Daily fraud reports: tells you if there was fraud yesterday
- Daily business reports: tells me how customers used AWS services yesterday

Real-time Big Data

- CloudWatch metrics**: what just went wrong *now***
- Real-time spending alerts/caps: **guaranteeing you can't overspend**
- Real-time analysis: **tells you what to offer the current customer *now***
- Real-time detection: **blocks fraudulent use *now***
- Fast ETL into Amazon Redshift: **how are customers using AWS services *now***

The Customer View

Developers View on Streaming Data Processing

Foundational Real-time Scenarios in Industry Segments

Scenarios	1 Accelerated Log/ Data Feed Ingest-Transform-Load	2 Continual Metrics/KPI Extraction	3 Real Time Data Analytics	4 Complex Stream Processing
Data Types	**IT infrastructure / Applications logs, Social media, Financial / Market data, Web Clickstream, Sensor data, Geo/Location data**			
Software/ Technology	IT server logs ingestion	IT operational metrics dashboards	Devices / Sensor Operational Intelligence	
Digital Ad Tech./ Marketing	Advertising Data aggregation	Advertising metrics like coverage, yield, conversion	Analytics on User engagement with Ads	Optimized bid/ buy engines
Financial Services	Market/ Financial Transaction order data collection	Financial market data metrics	Fraud monitoring, and Value-at-Risk assessment	Auditing of market order data
Consumer E-Commerce	Online customer engagement data aggregation	Consumer engagement metrics like page views, CTR	Customer clickstream analytics	Recommendation engines

Foundations for Streaming Data Processing

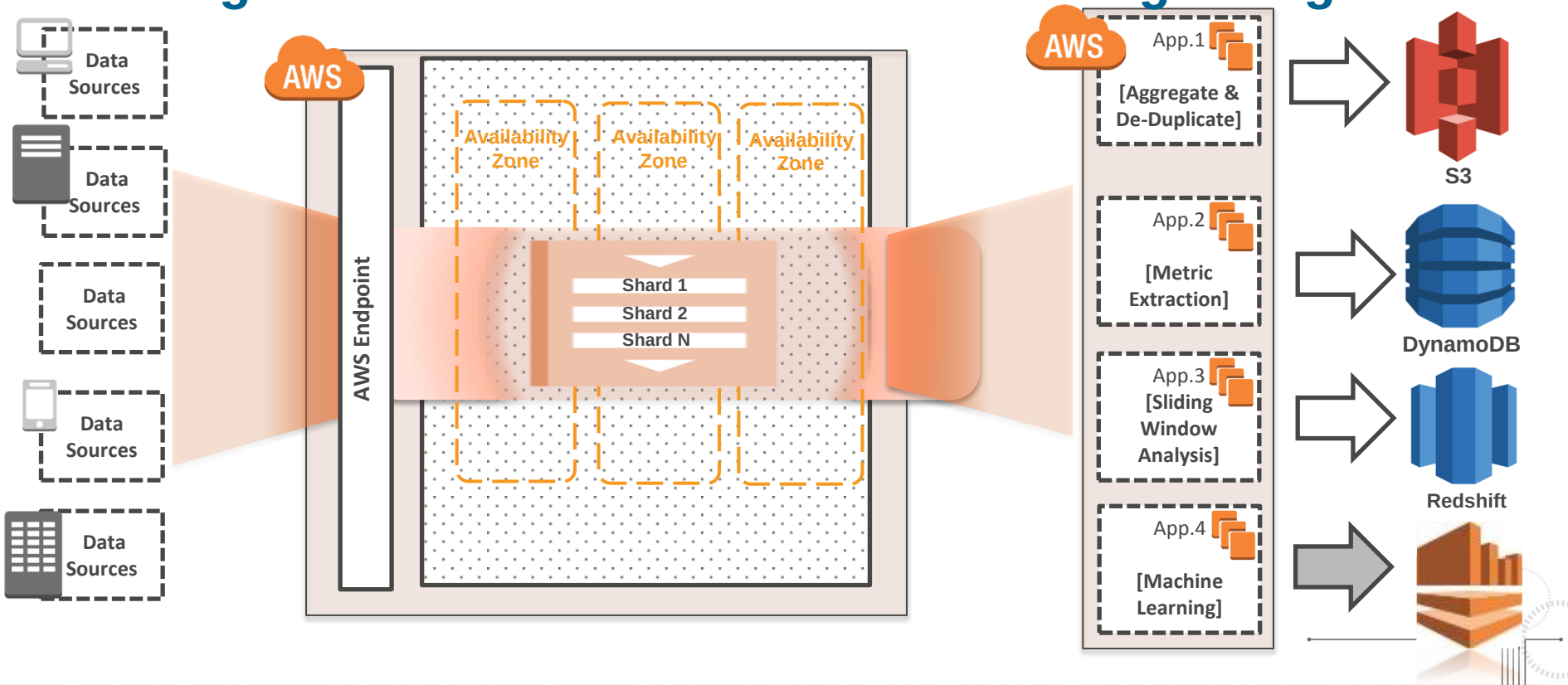
Learning from our customers

	Real-time Big Data Processing Wish list	Service Requirement
	Drive overall latencies of a few seconds, compared to minutes with typical batch processing	**Low end-to-end latency from data ingestion to processing**
	Scale up data ingestion to gigabytes per second, easily, without loss of durability	**Highly scalable, and durable**
	Scale up / down based on operational or business needs.	**Elastic**
	Offload complexity of load-balancing streaming data, distributed coordination services, and fault-tolerant data processing.	**Enable developers to focus on writing business logic for continual processing apps**
	Reduce operational burden of HW/ SW provisioning, patching, and operating a reliable real-time processing platform	**Managed service for real-time streaming data collection, processing and analysis.**

Amazon Kinesis

Introducing Amazon Kinesis

Managed Service for Real-Time Processing of Big Data

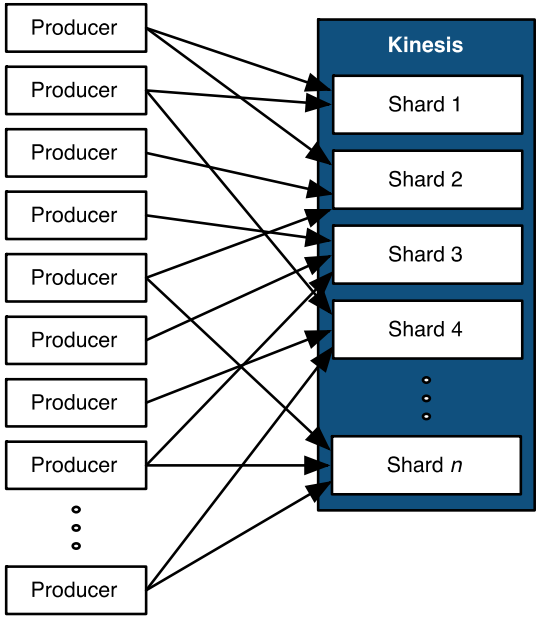


Putting data into Kinesis

Managed Service for Ingesting Fast Moving Data

- Streams are made of Shards
  - A Kinesis Stream is composed of multiple **Shards**
  - Each Shard ingests up to 1MB/sec of data and up to 1000 TPS
  - All data is stored for 24 hours
  - You scale Kinesis streams by adding or removing Shards

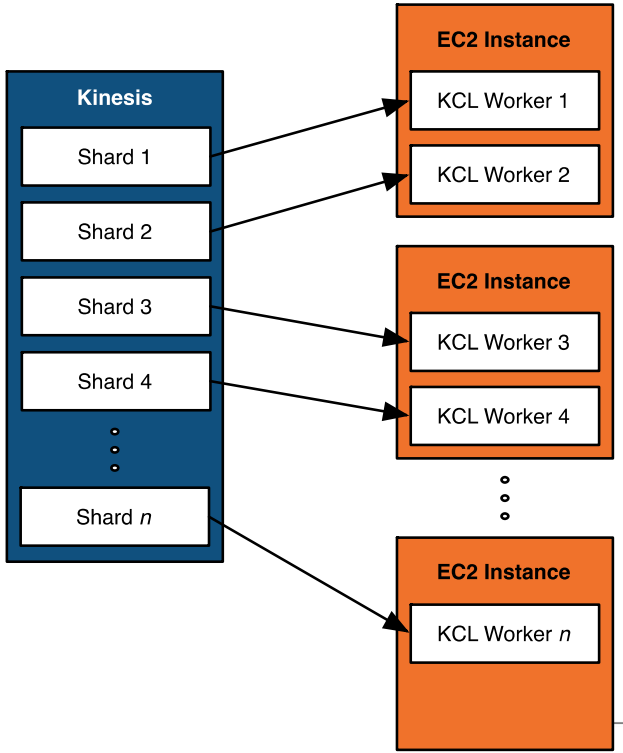
- Simple PUT interface to store data in Kinesis
  - Producers use a **PUT**call to store data in a Stream
  - A **Partition Key** is used to distribute the PUTs across Shards
  - A unique **Sequence #** is returned to the Producer upon a successful PUT call



Getting data out of Kinesis

Client library for fault-tolerant, at least-once, real-time processing

- In order to keep up with the stream, your application must:
  - Be distributed, to handle multiple shards
  - Be fault tolerant, to handle failures in hardware or software
  - Scale up and down as the number of shards increase or decrease
- Kinesis Client Library (KCL) helps with distributed processing:
  - Simplifies reading from the stream by abstracting your code from individual shards
  - Automatically starts a Kinesis Worker for each shard
  - Increases and decreases Kinesis Workers as number of shards changes
  - Uses checkpoints to keep track of a Worker's location in the stream
  - Restarts Workers if they fail
- Use the KCL with Auto Scaling Groups
  - Auto Scaling policies will restart EC2 instances if they fail
  - Automatically add EC2 instances when load increases
  - KCL will automatically redistribute Workers to use the new EC2 instances



Amazon Kinesis: Key Developer Benefits

Easy Administration

Managed service for real-time streaming data collection, processing and analysis. Simply create a new stream, set the desired level of capacity, and let the service handle the rest.

Real-time Performance

Perform continual processing on streaming big data. Processing latencies fall to a few seconds, compared with the minutes or hours associated with batch processing.

High Throughput. Elastic

Seamlessly scale to match your data throughput rate and volume. You can easily scale up to gigabytes per second. The service will scale up or down based on your operational or business needs.

S3, Redshift, & DynamoDB Integration

Reliably collect, process, and transform all of your data in real-time & deliver to AWS data stores of choice, with Connectors for S3, Redshift, and DynamoDB.

Build Real-time Applications

Client libraries that enable developers to design and operate real-time streaming data processing applications.

Low Cost

Cost-efficient for workloads of any scale. You can get started by provisioning a small stream, and pay low hourly rates only for what you use.

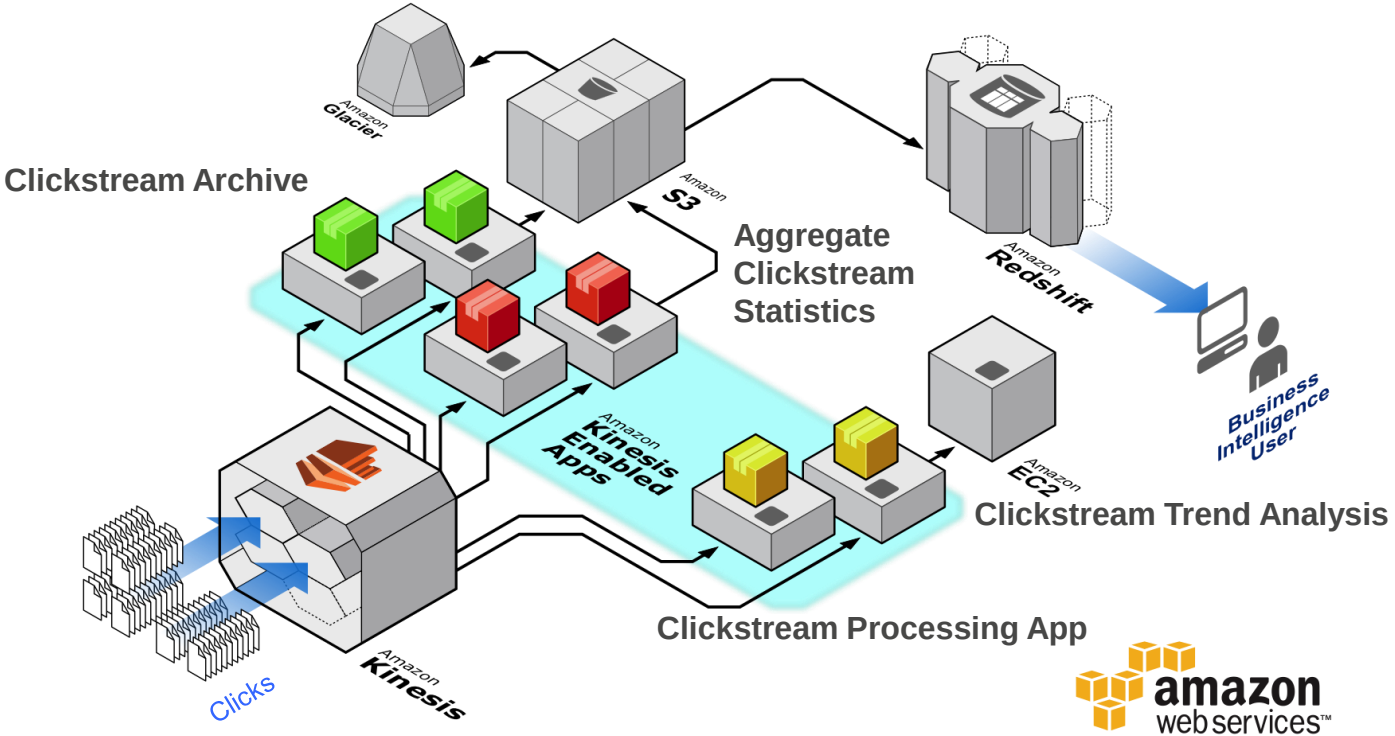
Sample Use Cases

Sample Customers using Amazon Kinesis (private beta)

Streaming big data processing in action

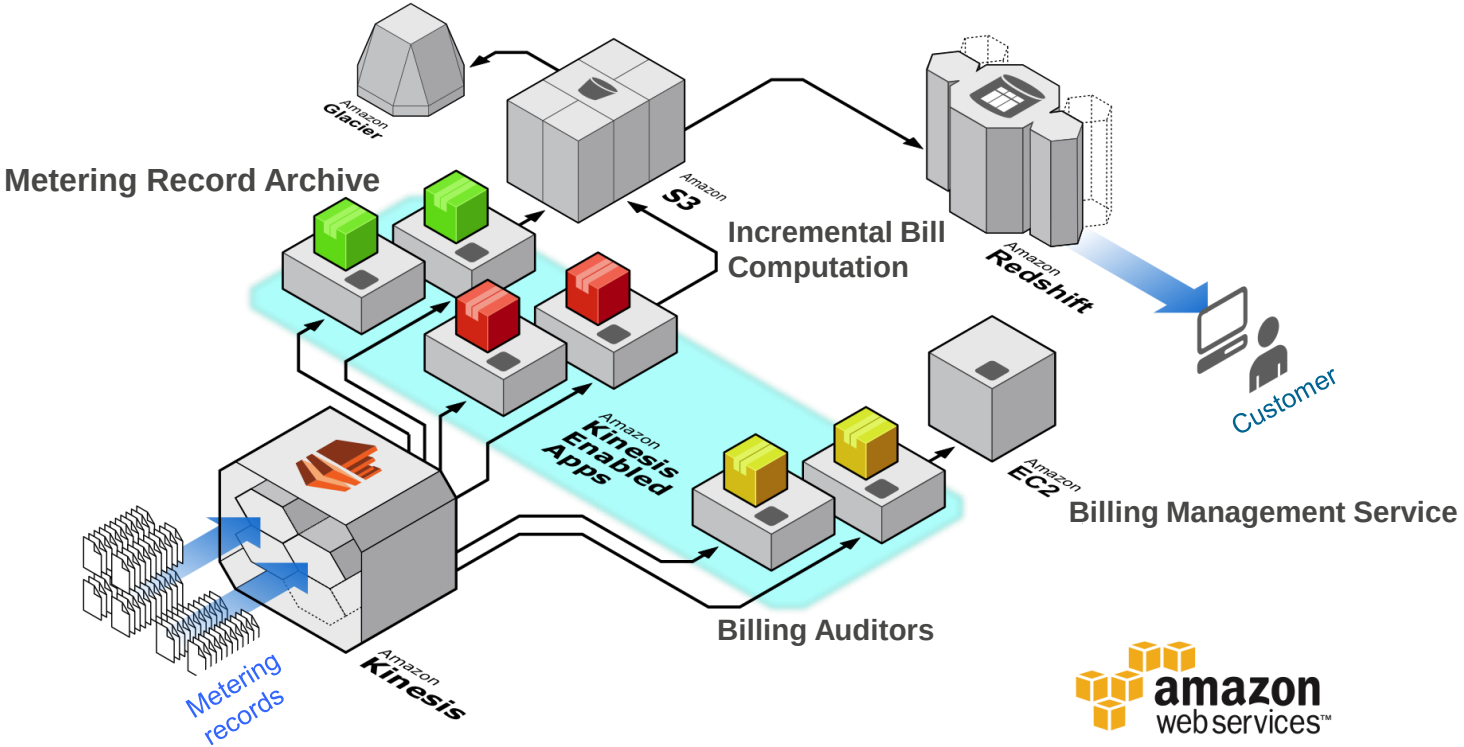
	Financial Services Leader	Digital Advertising Tech. Pioneer
	Maintain real-time audit trail of every single market/ exchange order	Generate real-time metrics, KPIs for online ads performance for advertisers
	Custom-built solutions operationally complex to manage, & not scalable	End-of-day Hadoop based processing pipeline slow, & cumbersome
	Kinesis enables customer to ingest all market order data reliably, and build real-time auditing applications	Kinesis enables customers to move from periodic batch processing to continual, real-time metrics and reports generation
	Accelerates time to market of elastic, real-time applications – while minimizing operational overhead	Generates freshest analytics on advertiser performance to optimize marketing spend, and increases responsive to clients

Clickstream Analytics with Amazon Kinesis



amazon web services™

Simple Metering & Billing with Amazon Kinesis



amazon web services™

The AWS Big Data Portfolio

COLLECT | STORE | ANALYZE | SHARE

Direct Connect

S3

EMR

EC2

S3

Import Export

DynamoDB

Redshift

Data Pipeline

Glacier

Kinesis

Please Attend BDT311

Level 300 talk by Marvin Theimer, Distinguished Engineer

- San Polo 3501A – Friday at 11:30 AM
- Amazon Kinesis core concepts deep dive
- Overview of a sample Kinesis application

AWS re:Invent

Please give us your feedback on this presentation

BDT 103

As a thank you, we will select prize winners daily for completed surveys!

amazon web services