Real Time Big Data Processing
Global Infrastructure

- US-WEST (Oregon)
- EU-WEST (Ireland)
- ASIA PAC (Tokyo)
- ASIA PAC (Sydney)
- ASIA PAC (Singapore)
- GOV CLOUD
- US-EAST (Virginia)
- US-WEST (San Francisco)
- SOUTH AMERICA (Sao Paulo)
Ingestion  |  Integration
### Elastic Block Store

- **High performance block storage device**
- **1GB to 1TB in size**
- **Mount as drives to instances with snapshot/cloning functionalities**

- **Availability**: 99.99%
- **Durability**: 99.999999999%

- **Is a Web Store**
  - Not a file system
  - No Single Points of Failure
  - Eventually consistent

### Simple Storage Service

- **Highly scalable object storage for the internet**
  - 1 byte to 5TB in size
  - **99.999999999% durability**

#### Table

<table>
<thead>
<tr>
<th><strong>Paradigm</strong></th>
<th>Object store</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance</strong></td>
<td>Very Fast</td>
</tr>
<tr>
<td><strong>Redundancy</strong></td>
<td>Across Availability Zones</td>
</tr>
<tr>
<td><strong>Security</strong></td>
<td>Public Key / Private Key</td>
</tr>
<tr>
<td><strong>Pricing</strong></td>
<td>$0.095/GB/month</td>
</tr>
<tr>
<td><strong>Typical use case</strong></td>
<td>Write once, read many</td>
</tr>
<tr>
<td><strong>Limits</strong></td>
<td>100 Buckets, Unlimited Storage, 5TB Objects</td>
</tr>
</tbody>
</table>
Peak Requests: 1.2 Million / Second

Objects in S3

Billions

Q4 2007 | Q4 2008 | Q4 2009 | Q4 2010 | Q4 2011 | Q4 2012 | Today

14 | 40 | 102 | 262 | 762 | 1300 | 2100
Storage Lifecycle Integration

Simple Storage Service

Highly scalable object storage
1 byte to 5TB in size
99.999999999% durability

Glacier

Long term object archive
Extremely low cost per gigabyte
99.999999999% durability
Complex Analytics

Unstructured Data

Parallel ETL
Elastic MapReduce

Managed, elastic Hadoop (1.x & 2.x) cluster
Integrates with S3, DynamoDB and Redshift
Install Storm, Spark & Shark, Hive, Pig, Impala & End User Tools Automatically
Support for Spot Instances
Integrated HBase NOSQL Database
Analytics Orchestration

Data Pipeline

 Automatically Provision EC2 & EMR Resources
 Manage Dependencies & Scheduling
 Automatically Retry and Notify of Success & Failure
Structured Data Management
DynamoDB

Provisioned throughput NoSQL database
Fast, predictable, configurable performance
Fully distributed, fault tolerant HA architecture
Integration with EMR & Hive
Database

Redshift

Managed Massively Parallel Petabyte Scale Data Warehouse
Streaming Backup/Restore to S3
Load data from S3, DynamoDB and EMR
Extensive Security Features
Scale from 160 GB -> 1.6 PB Online
Amazon Redshift parallelizes and distributes everything

Query
Load
Backup
Restore
Resize

Common BI Tools

JDBC/ODBC

Leader Node

Compute Node

10GigE Mesh

Compute Node

Compute Node
VOLUME

VELOCITY

VARIETY
Big Data is Moving Fast…

- IT Application logs, Infrastructure logs, Metering, Audit logs, Change and Configuration logs
- Web sites / Mobile Apps/ Ads Clickstream, User Engagement
- Sensor data Weather, Smart Grids, Wearables
- Social Media, User Content 450MM+ Tweets/day
The Move to Real Time

Analytics Recency Requirements are moving from Daily to Minute Granularity

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
Amazon Kinesis

Managed Service for Real Time Big Data Processing
Create Streams to Produce & Consume Data
Elastically Add and Remove Shards for Performance
Use Kinesis Worker Library to Process Data
Integration with S3, Redshift and Dynamo DB
Amazon Kinesis

Data Sources

AWS Endpoint

Availability Zone
Availability Zone
Availability Zone

Shard 1
Shard 2
Shard N

App.1
[Aggregate & De-Duplicate]

App.2
[Metric Extraction]

App.3
[Sliding Window Analysis]

App.4
[Machine Learning]

S3

DynamoDB

Redshift

Availability Zone
Availability Zone
Availability Zone
Kinesis Connectors

Analytics Tooling Integration

- S3
  - Batch Write Files for Archive into S3
  - Sequence Based File Naming
- Redshift
  - Once Written to S3, Load to Redshift
  - Manifest Support
  - User Defined Transformers
- DynamoDB
  - BatchPut Append to Table
  - User Defined Transformers
- Storm
  - Use Kinesis as a Spout
- Provided as Open Source on GitHub
  - amazon-kinesis-connectors
EMR Integration with Kinesis

- Read Data Directly into Hive, Pig, Streaming and Cascading from Kinesis Streams
- No Intermediate Data Persistence Required

- Simple way to introduce real time sources into Batch Oriented Systems
- Multi-Application Support & Automatic Checkpointing
Integrated Real Time Analytics
Thank you!

- Join the sessions in our AWS Lab
- Meet the team at our Amazon Web Services Partner Village