Big Data on AWS

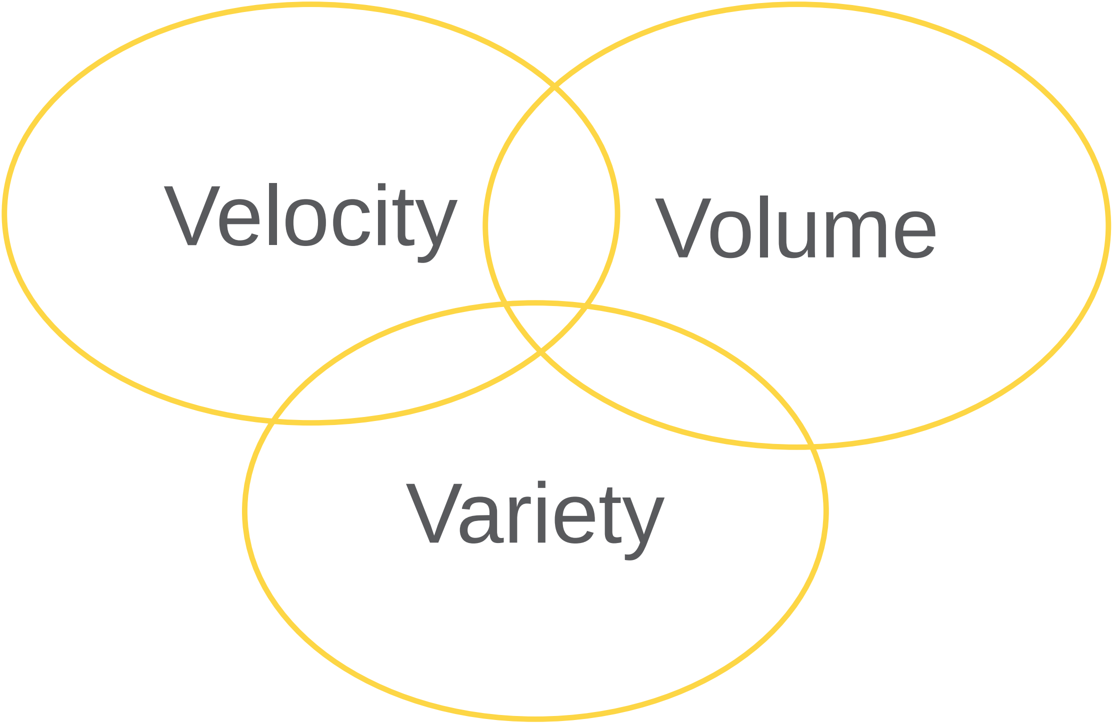
Big Data Agility and Performance Delivered in the Cloud

© 2015, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Big Data

Technologies and techniques for working productively with massive amounts of data at any scale in either batch or real-time.

Three Vs of Big Data

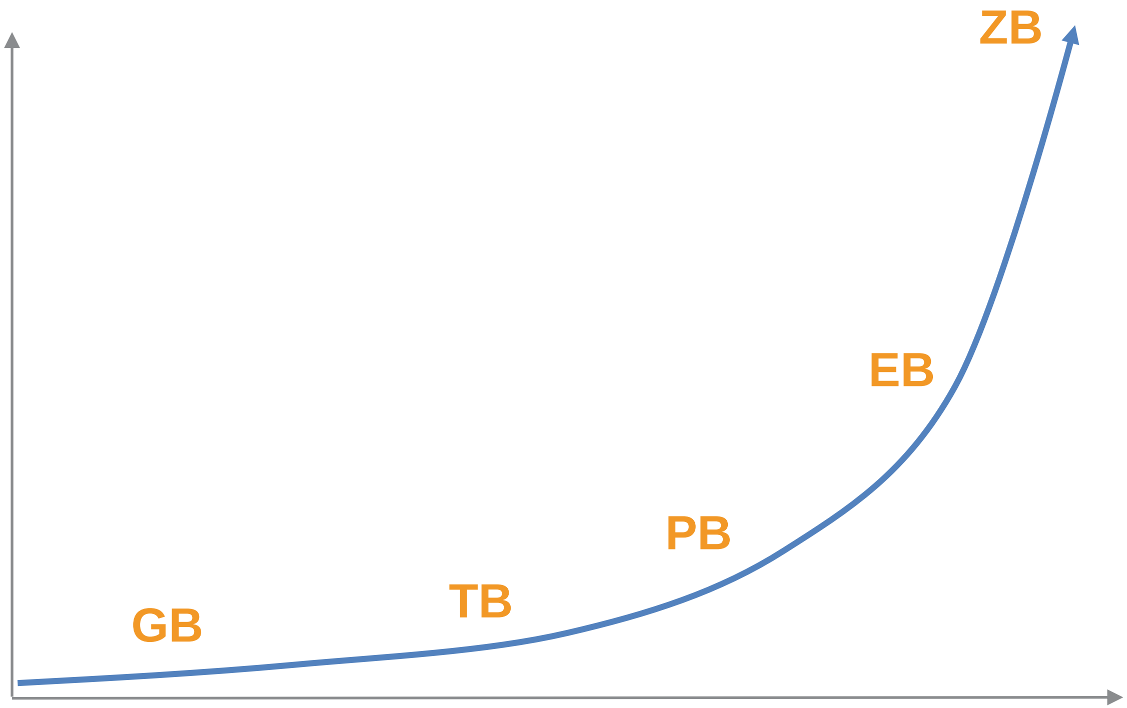


Velocity

Volume

Variety

Big Data: Unconstrained Growth



Unstructured data growth is explosive

95% of the 1.2 zettabytes of data in the digital universe is unstructured

Machine data and IoT will only steepen the curve

70% of this data is user-generated content

Source: IDC, The Internet of Things: Getting Ready to Embrace Its Impact on the Digital Economy, March 2016.

Big Data Sources

Sources

Relational

NoSQL

Web servers

Mobile phones/Tablets

3rd party feeds

IoT

Clickstream

Big Data Formats and Velocity

Formats

Structured

Unstructured

Text

Binary

Velocity

Real-time/Near Real-time

Batched

Managed Services for Analytics

Retrospective
analysis and
reporting

Here-and-now
real-time processing
and dashboards

Predictions
to enable smart
applications

Why Big Data?

Get answers faster and be able to ask questions not possible to today.

Security threat detection

User Behavior Analysis

Smart Application (Machine Learning)

Business Intelligence

Fraud detection

Financial Modeling and Forecasting

Spending optimization

Real-time alerting

Elastic and highly scalable

+

No upfront capital expense

+

Only pay for what you use

+

Available on-demand

= **the Cloud removes constraints**

The Cloud Was Built for Big Data

Agility: Try more, fail fast, go big or start small, and process data at any scale

Broadest and Deepest Capabilities: Access 70+ managed Big Data services to address any workload

Get to Insights Faster: Focus on data science not the heavy undifferentiated lift of managing raw data

Scalability: Run jobs any time, without guessing capacity or limiting functionality

Low Cost: Pay only for the IT you use, when you use it

Data Migrations Made Easy: Move exabyte-scale data to the cloud quickly and cost-effectively

Big Data was Meant for the Cloud

Big Data

Variety, volume, and velocity requiring new tools

Iterative, experimental style of data manipulation and analysis

Potentially massive datasets

Absolute performance not as critical as “time to results”; shared resources are a bottleneck

Frequently non-steady-state workloads with peaks and valleys

Cloud Computing

Variety of compute, storage, and networking options

Iterative, experimental style of IT infrastructure deployment and usage

Parallel compute projects allow each workgroup to have more autonomy and get faster results

At its most efficient with highly variable workloads

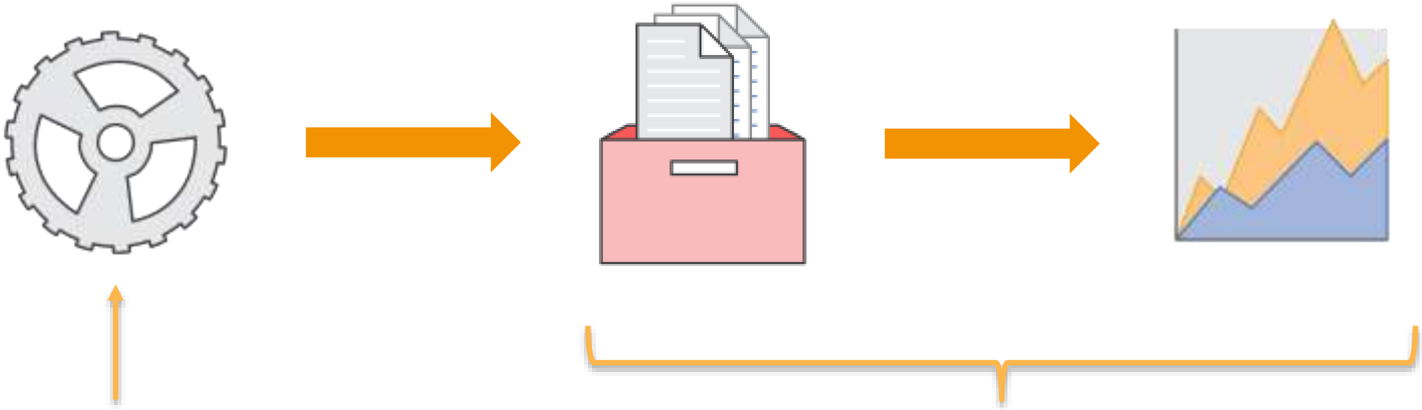
Massive, virtually unlimited capacity

Common Big Data Flow

Generate

Collect, Orchestrate, Store

Analyze



Lower Cost
Increased Velocity

Traditionally - Highly constrained

AWS Big Data Platform

Collect

- Direct Connect
- Import Export
- AWS Snowball
- AWS IoT
- Kinesis
- AWS Database Migration Service

Orchestrate

- AWS Lambda
- AWS Data Pipeline
- Amazon SNS
- Amazon SWF
- AWS Glue

Store

- S3
- Glacier
- DynamoDB
- Amazon Aurora

Analyze

- EMR
- EC2
- Redshift
- Machine Learning
- Amazon Kinesis
- Amazon QuickSight
- Amazon Athena

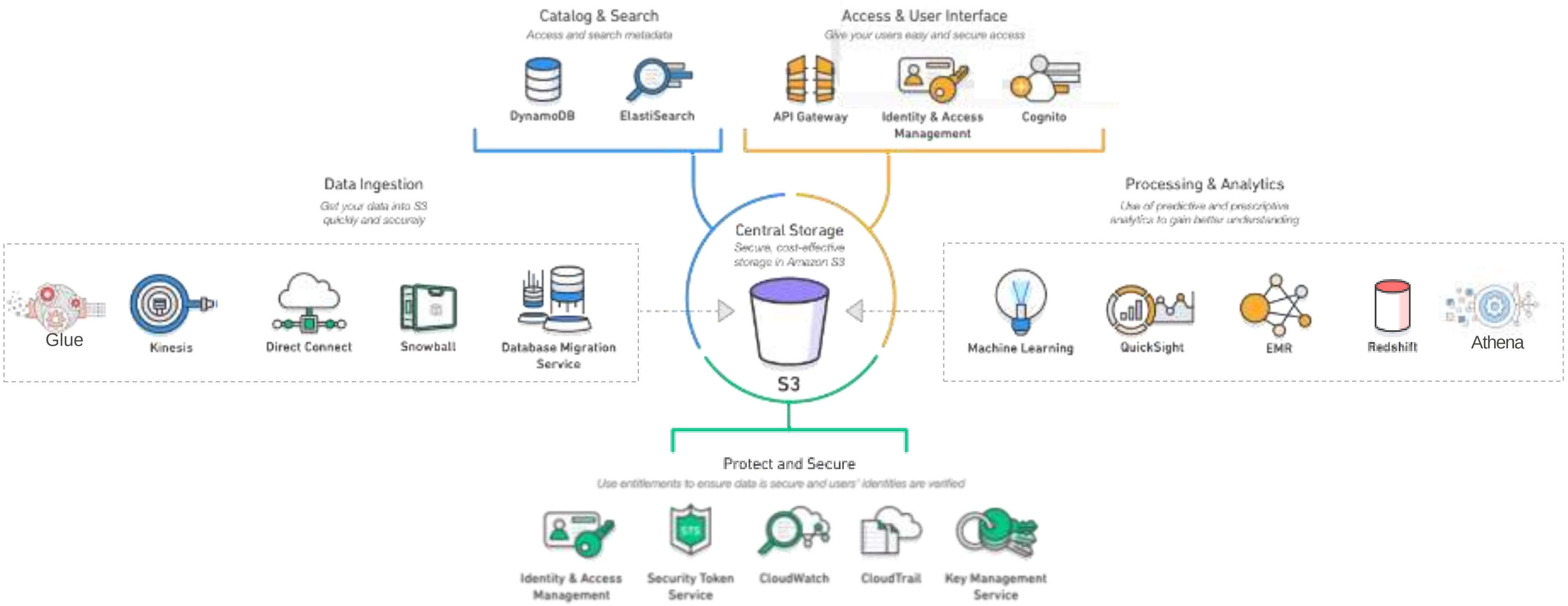
amazon web services

No one tool rules them all

The AWS Approach

- Flexible - Use the best tool for the job
  - Data structure, latency, throughput, access patterns
- Low Cost - Big data ≠ big cost
- Scalable – Data should be immutable (append-only)
  - Batch/speed/serving layer
- Minimize Admin Overhead - Leverage AWS managed services
  - No or very low admin
- Be Agile – Fail fast, test more, optimize Big Data at a lower cost

Sample Reference Architecture: Data Lake



Catalog & Search
Access and search metadata
DynamoDB
ElastiSearch
Access & User Interface
Give your users easy and secure access
API Gateway
Identity & Access Management
Cognito
Data Ingestion
Get your data into S3 quickly and securely
Glue
Kinesis
Direct Connect
Snowball
Database Migration Service
Central Storage
Secure, cost-effective storage in Amazon S3
S3
Processing & Analytics
Use of predictive and prescriptive analytics to gain better understanding
Machine Learning
QuickSight
EMR
Redshift
Athena
Protect and Secure
Use entitlements to ensure data is secure and users' identities are verified
Identity & Access Management
Security Token Service
CloudWatch
CloudTrail
Key Management Service

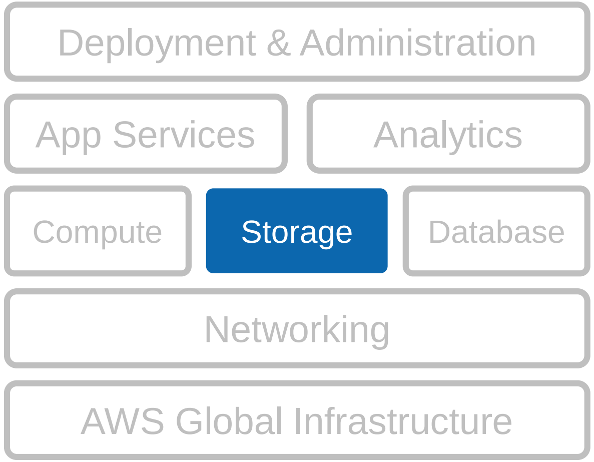
Summary

- Build sophisticated Big Data applications cost-effectively and support retrospective, real-time and predictive analysis
- You can build incrementally, scale automatically and add use cases as you go
- AWS delivers added benefits of security and auditing features to enable you to meet your stringent requirements
- Build hybrid applications that span across your datacenters and the AWS Cloud

AWS Big Data Services

© 2015, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Amazon S3



Scalable object storage for the Internet

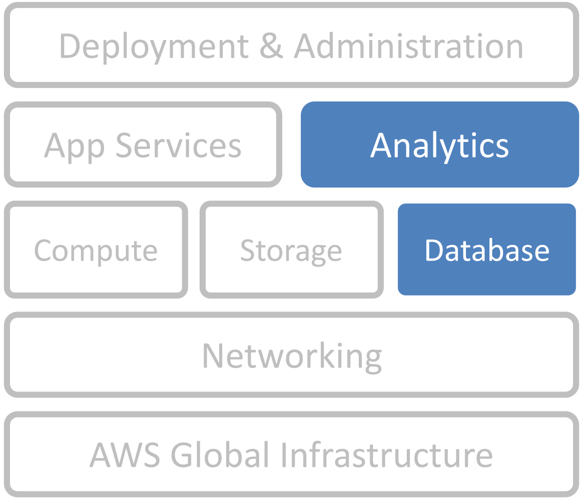
1 byte to 5 TB in size per object + unlimited number of objects

99.999999999% durability, 99.99% availability

Regional service, no single points of failure

Server side encryption

Amazon Redshift



Managed Massively Parallel Petabyte Scale Data Warehouse

Streaming Backup/Restore to S3

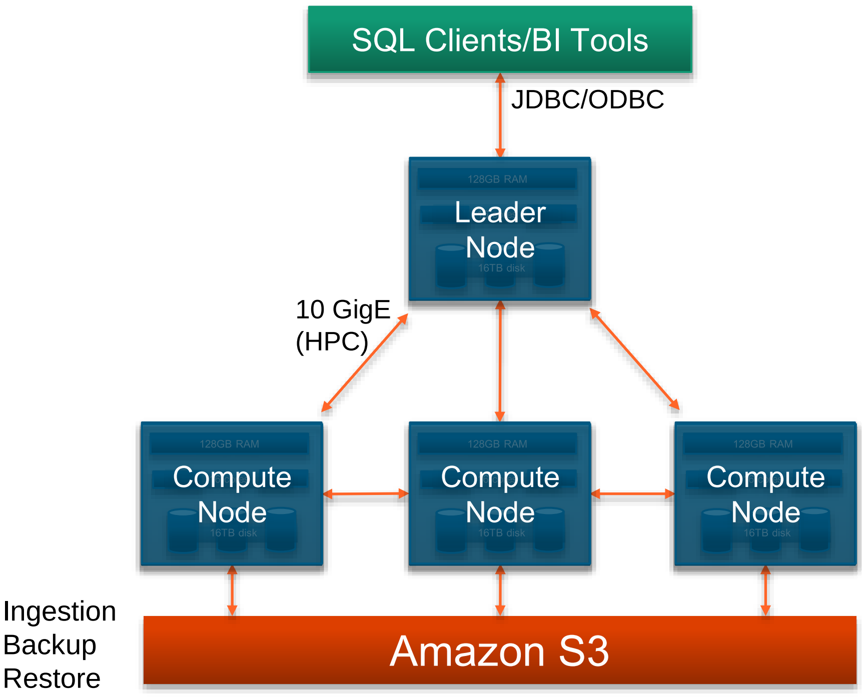
Load data from S3, DynamoDB and EMR

Extensive Security Features

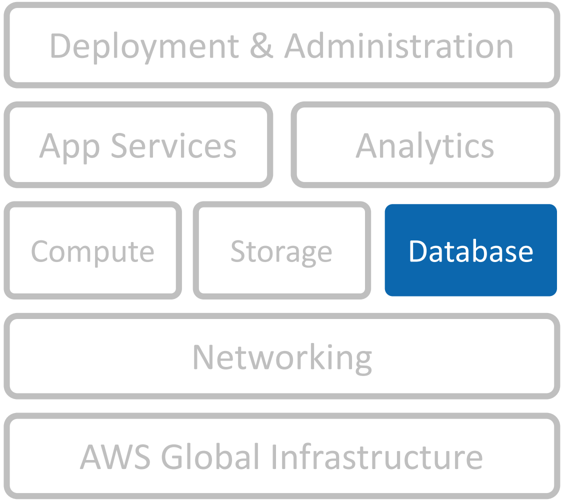
Online Scaling from 160 GB -> 2 PB

Amazon Redshift

- **Scalability & Elasticity**
  - Resize or scale - Number or type of nodes can be changed with a few clicks
- **Durability and Availability**
  - Replication
  - Backup
  - Automated recovery from failed drives & nodes
- **Interfaces**
  - JDBC/ODBC interface with BI/ETL tools
  - Amazon S3 or DynamoDB
- **Anti-patterns**
  - Small datasets (smallest database 160GB)
  - OLTP
  - Unstructured Data
  - Blob Data



Amazon DynamoDB



Deployment & Administration

App Services

Analytics

Compute

Storage

Database

Networking

AWS Global Infrastructure

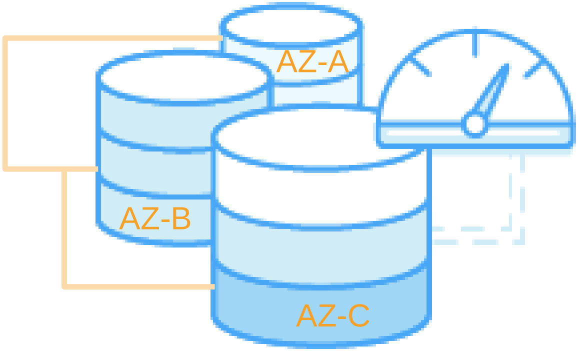
Fully managed NoSQL database

Single-Digit Millisecond latency at scale

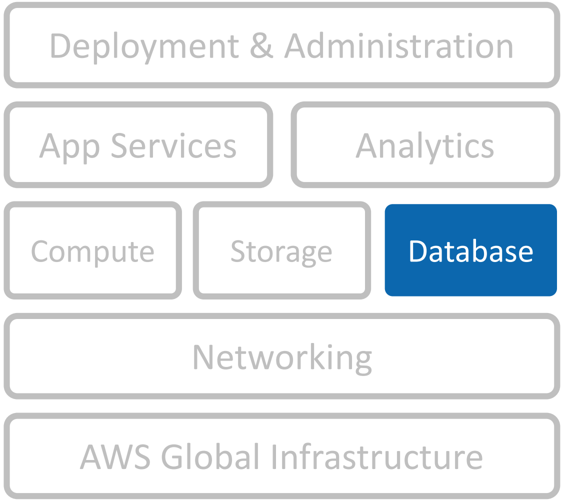
Supports document and key-value

Amazon DynamoDB

- **Durability and Availability**
  - Three Availability Zones (AZ)
- **Interfaces**
  - AWS Management Console
  - API's
  - SDK's
- **Anti-patterns**
  - Application tied to traditional relational database
  - Joins and or complex transactions
  - BLOB data
  - Large data with low I/O rate



Amazon Aurora

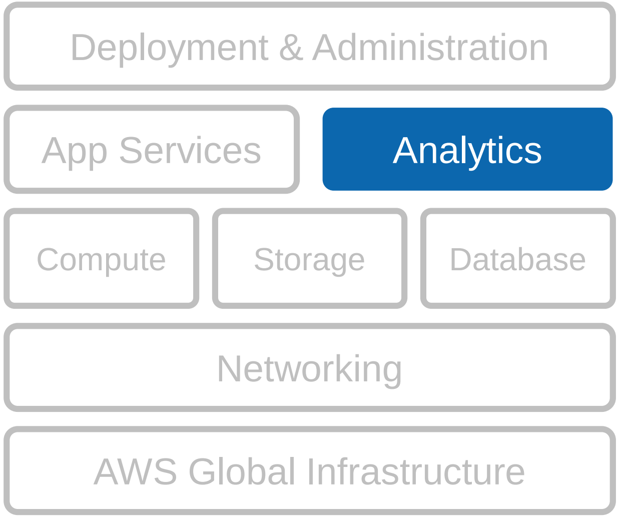


5x performance at 1/10th the cost of alternatives

Fully managed MySQL-compatible database

Fast with 500K reads/100K writes per second

Amazon Kinesis

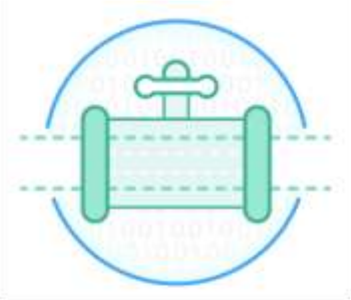


Ingest streaming data

Process data in real-time

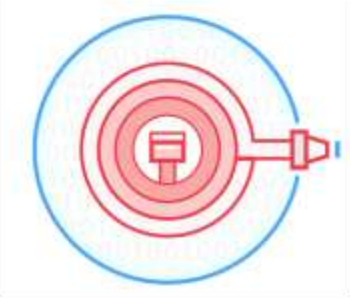
Store terabytes of data per hour

Amazon Kinesis



Amazon Kinesis Streams

Build your own custom applications that process or analyze streaming data



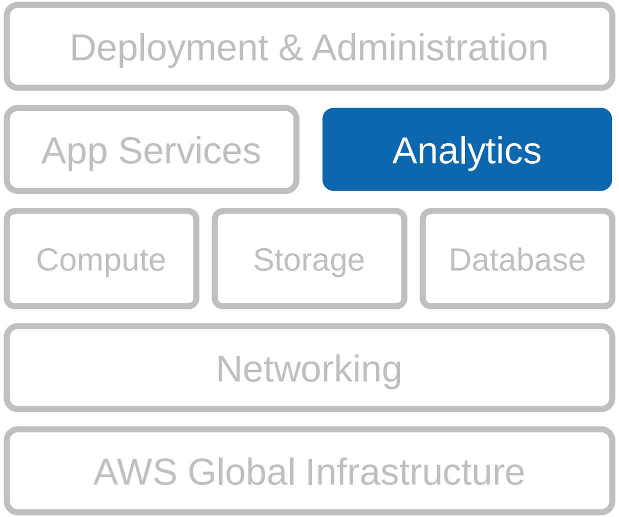
Amazon Kinesis Firehose

Easily load massive volumes of streaming data into Amazon S3 and Redshift

Amazon Kinesis Analytics

Easily analyze data streams using standard SQL queries

Amazon EMR



Deployment & Administration

App Services

Analytics

Compute

Storage

Database

Networking

AWS Global Infrastructure

Scalable Hadoop/Spark clusters as a service

Launch a cluster in minutes

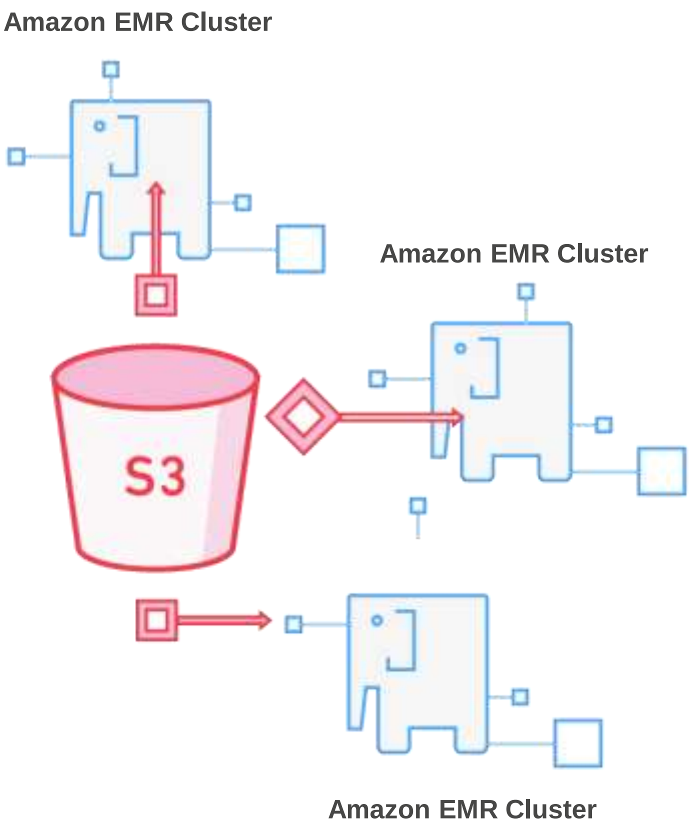
Hadoop, Hive, Spark, Presto, HBase, etc.

Easy to use; fully managed

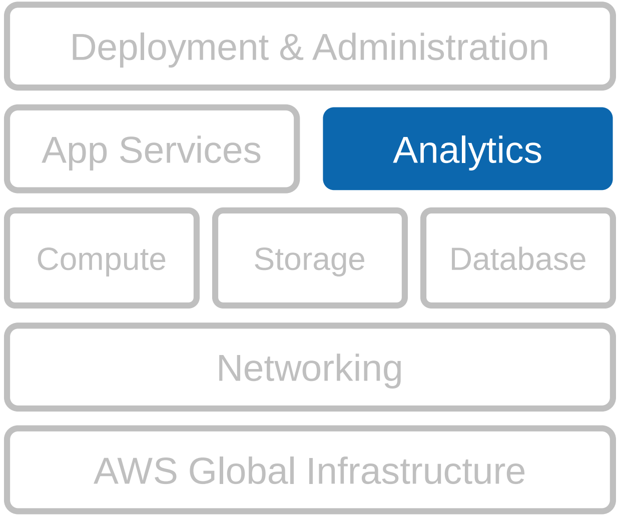
HDFS, Amazon EBS, and S3 file systems

Amazon EMR

- **Scalability & Elasticity**
  - Resize a running cluster based on how much work is needed to be done.
- **Durability and Availability**
  - Fault tolerant for slave node (HDFS)
  - Backup to S3 for resilience against master node failures
- **Standard Interfaces**
  - Hive, Pig, Spark, Hbase, Impala, Hunk, Presto, other popular tools



Amazon QuickSight



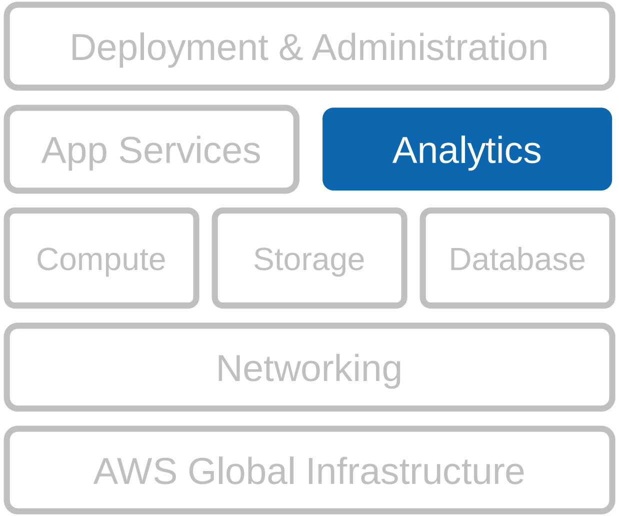
BI service performs ad-hoc analysis

Build visualizations

Share and collaborate via storyboards

Native access on major mobile platforms

Machine and Deep Learning



Deployment & Administration

App Services

Analytics

Compute

Storage

Database

Networking

AWS Global Infrastructure

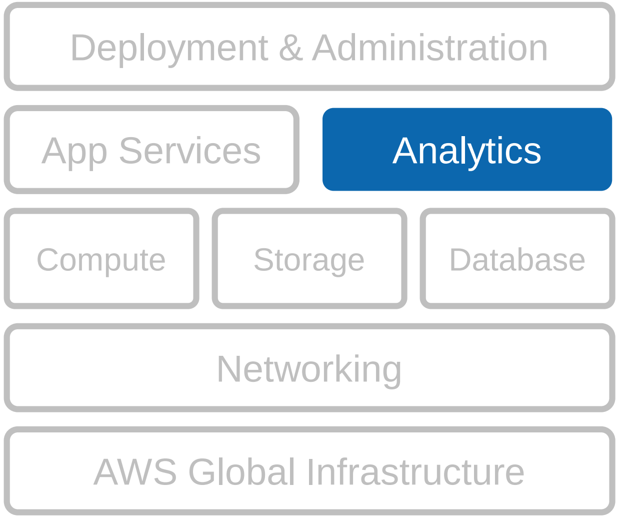
Amazon Machine Learning
scalable and robust implementations of industry-standard ML supervised learning algorithms

Amazon Lex
Conversational interfaces through Voice or Text
Backend powering Alexa

Amazon Polly
Cloud Native TTS (Text to Speech)
47 lifelike voices/24 languages (on growing)
Low-latency for real-time applications

Amazon Rekognition
Deep learning-based image recognition
Object/Scene detection, facial analysis and comparison

Amazon Elasticsearch Service



Deployment & Administration

App Services

Analytics

Compute

Storage

Database

Networking

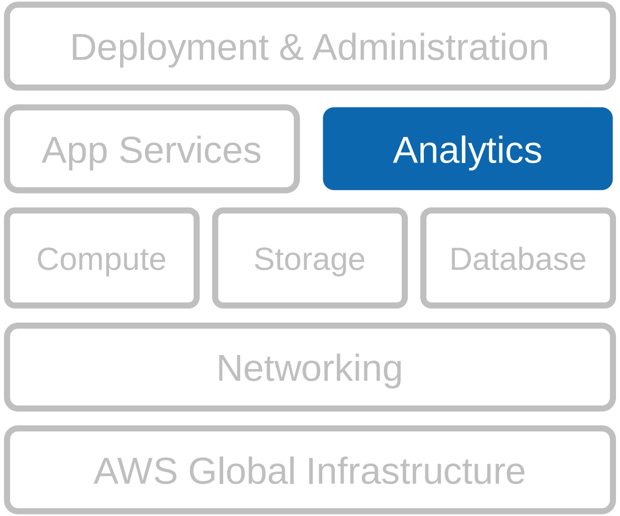
AWS Global Infrastructure

Setup Elasticsearch cluster in minutes

Integrated with Logstash and Kibana

Scale Elasticsearch clusters seamlessly

Amazon Athena



Deployment & Administration

App Services

Analytics

Compute

Storage

Database

Networking

AWS Global Infrastructure

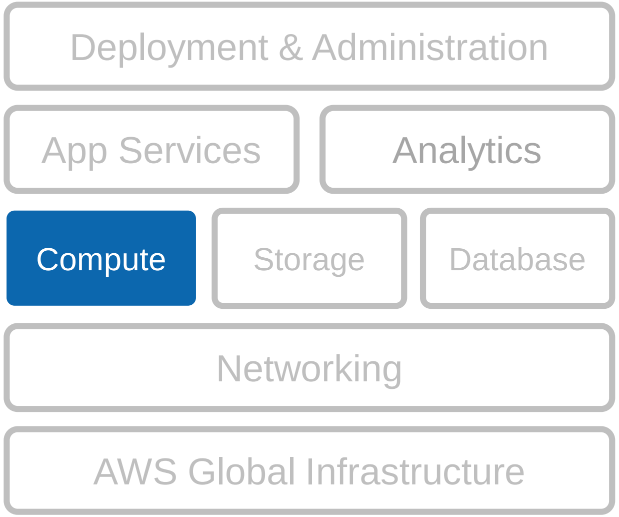
Query and analyze Amazon S3 data with standard (ANSI) SQL

No ETL required

Serverless and simple

Pay only for queries you run

Amazon EC2



Deployment & Administration

App Services

Analytics

Compute

Storage

Database

Networking

AWS Global Infrastructure

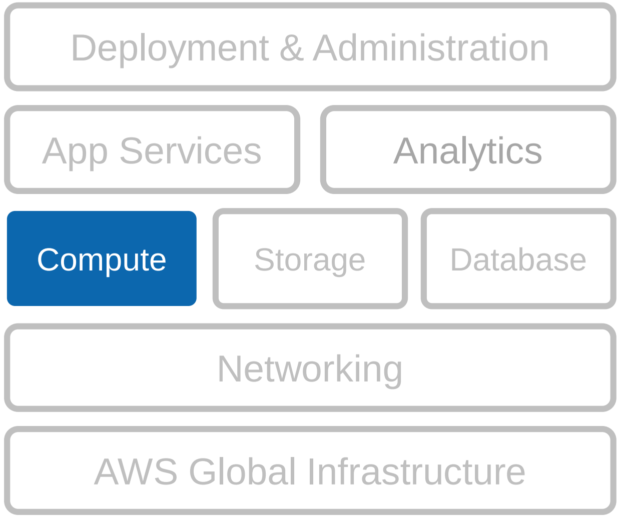
Scale up or down as needed

Pay for what you use

Largest select of instance types

Do-it-yourself big data applications

AWS Lambda



Deployment & Administration

App Services

Analytics

Compute

Storage

Database

Networking

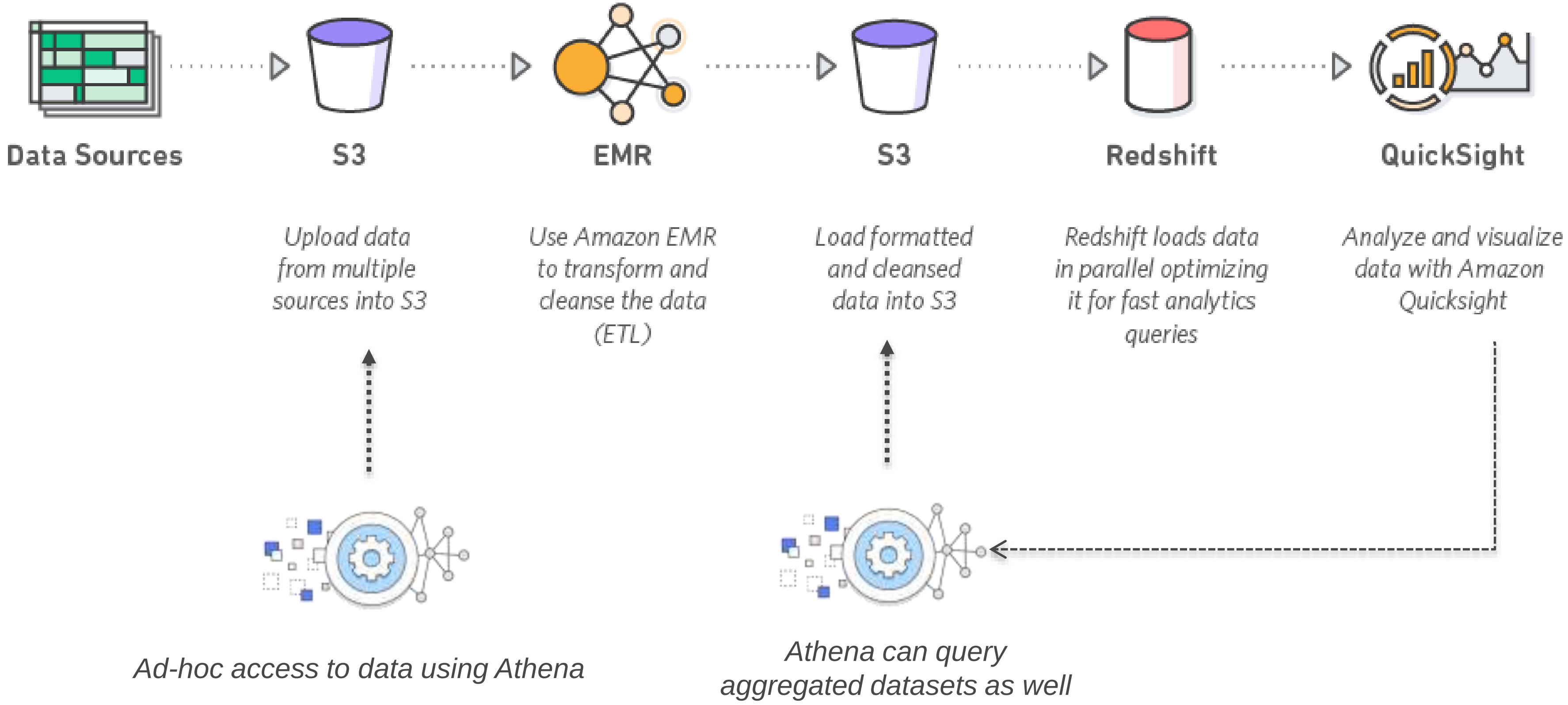
AWS Global Infrastructure

Event driven, fully managed compute

No Infrastructure to Manage

Automatic Scaling

A Sample Batch Analytics Pipeline



Data Sources

S3

Upload data from multiple sources into S3

EMR

Use Amazon EMR to transform and cleanse the data (ETL)

S3

Load formatted and cleansed data into S3

Redshift

Redshift loads data in parallel optimizing it for fast analytics queries

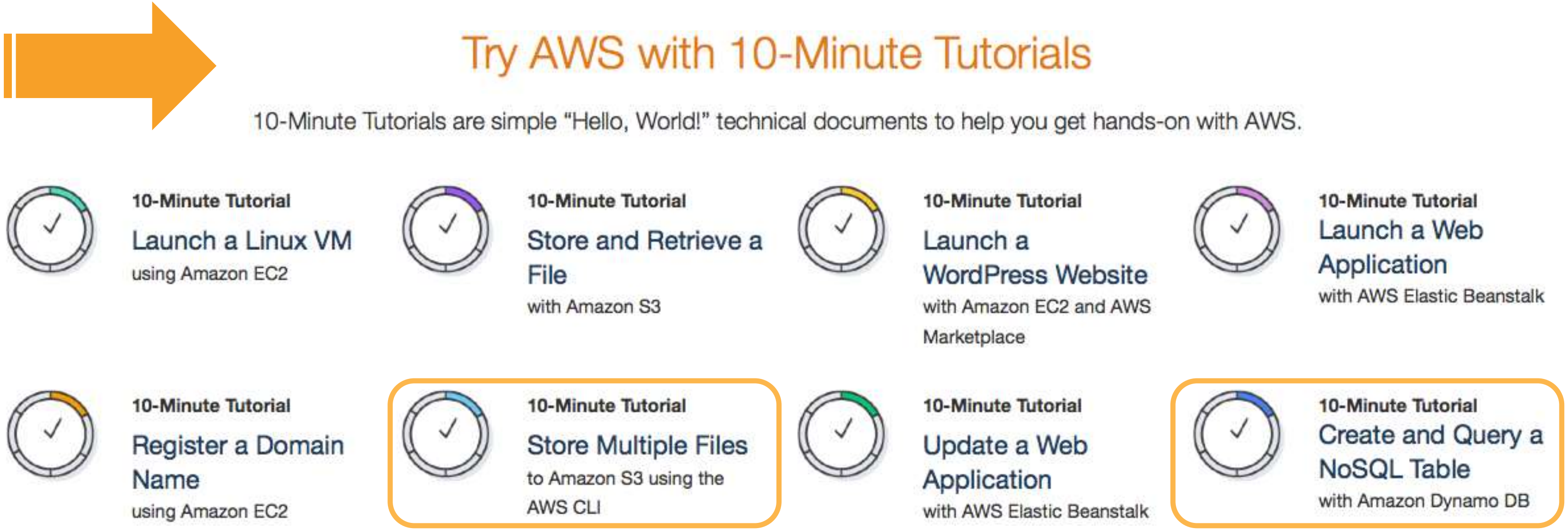
QuickSight

Analyze and visualize data with Amazon Quicksight

Ad-hoc access to data using Athena

Athena can query aggregated datasets as well

Getting Started: Tutorials & Blog



Try AWS with 10-Minute Tutorials

10-Minute Tutorials are simple “Hello, World!” technical documents to help you get hands-on with AWS.

10-Minute Tutorial
Launch a Linux VM
using Amazon EC2

10-Minute Tutorial
Store and Retrieve a File
with Amazon S3

10-Minute Tutorial
Launch a WordPress Website
with Amazon EC2 and AWS Marketplace

10-Minute Tutorial
Launch a Web Application
with AWS Elastic Beanstalk

10-Minute Tutorial
Register a Domain Name
using Amazon EC2

10-Minute Tutorial
Store Multiple Files
to Amazon S3 using the AWS CLI

10-Minute Tutorial
Update a Web Application
with AWS Elastic Beanstalk

10-Minute Tutorial
Create and Query a NoSQL Table
with Amazon Dynamo DB

Subscribe to the AWS Big Data Blog: http://blogs.aws.amazon.com/bigdata/

Customer Success.

Powered by AWS.

FINRA Analyzes Billions of Transactions Daily

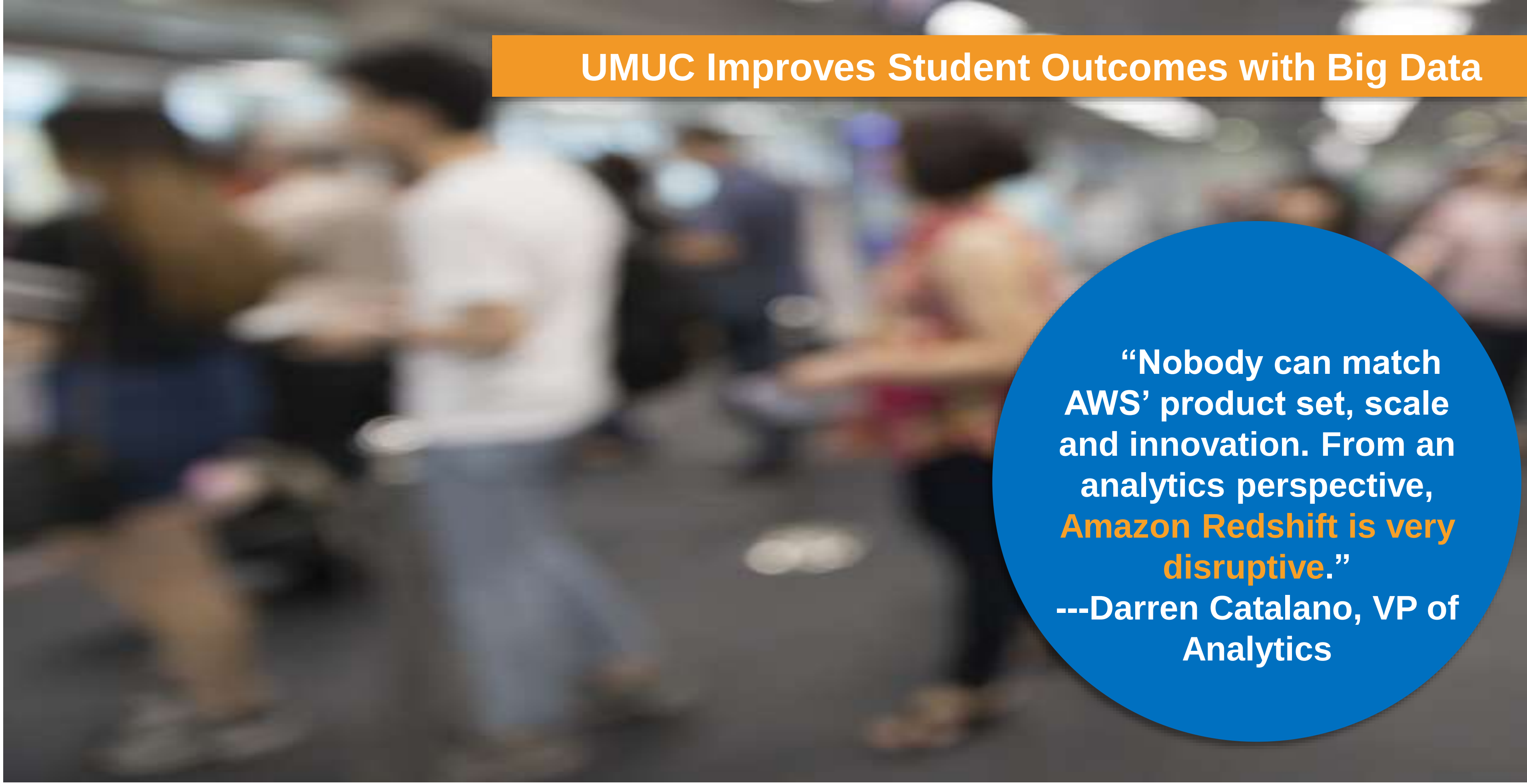
To respond to rapidly changing market dynamics, FINRA moved **75% of its operations to Amazon Web Services**, using AWS to analyze **75B records a day**.

Just Giving Creates a Big Data Platform on AWS

"Before AWS, [we were] basing decisions on a single high-level data source. **Now we can extract much more granular data** based on millions of donations…and **use that information to provide a better platform for our visitors**."

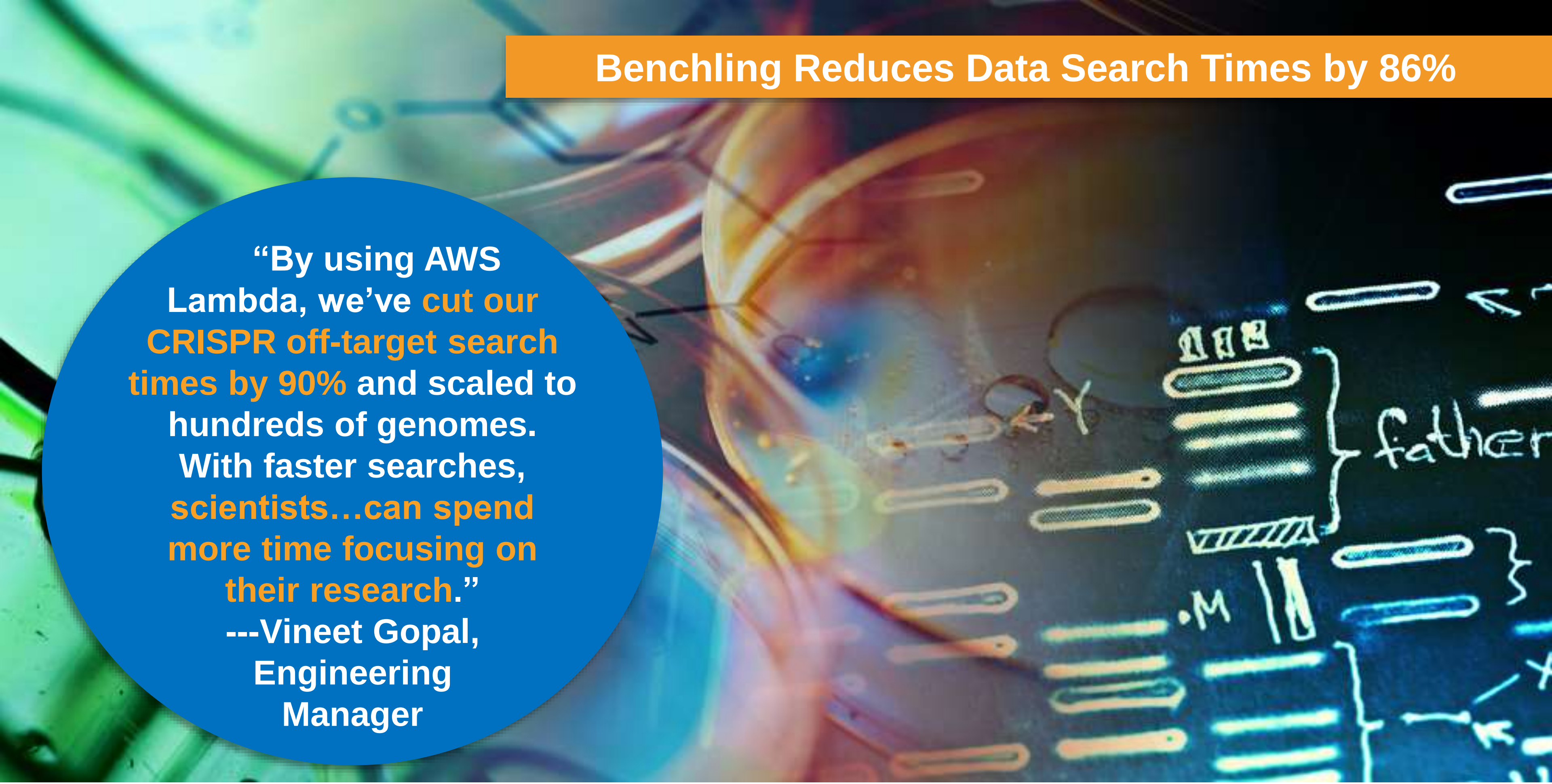
-Richard Atkinson, CIO

UMUC Improves Student Outcomes with Big Data



“Nobody can match AWS’ product set, scale and innovation. From an analytics perspective, **Amazon Redshift is very disruptive**.”
---Darren Catalano, VP of Analytics

Benchling Reduces Data Search Times by 86%

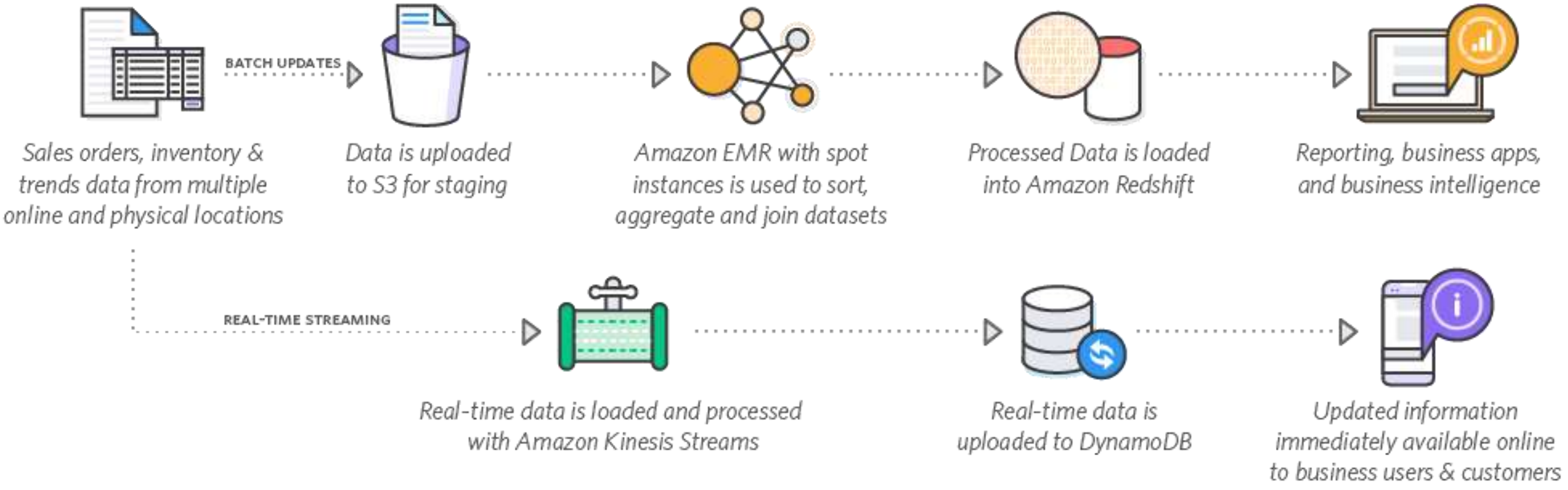


"By using AWS Lambda, we've **cut our CRISPR off-target search times by 90%** and scaled to hundreds of genomes. With faster searches, **scientists…can spend more time focusing on their research.**"
---Vineet Gopal, Engineering Manager

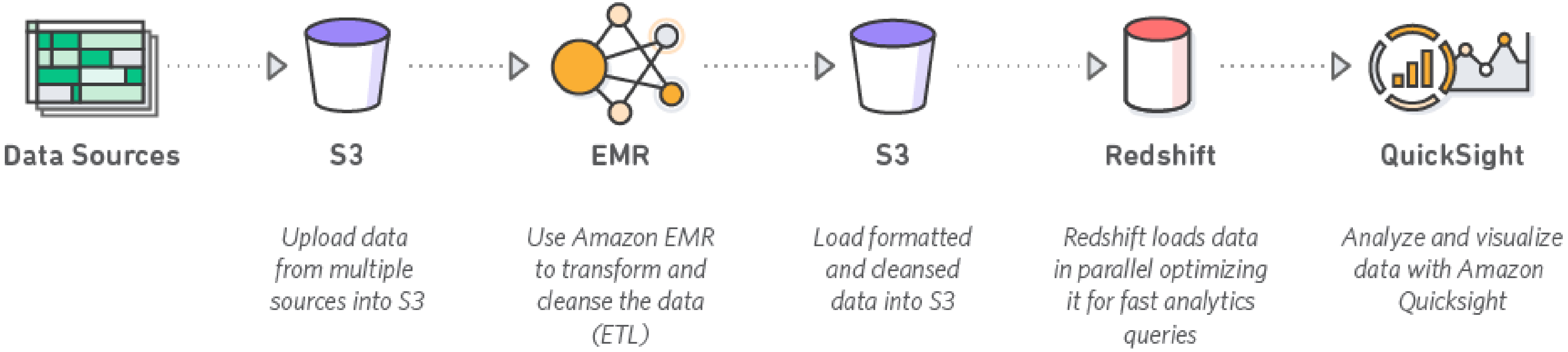
Optional Slides

© 2015, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

On-demand Big Data Analytics


BATCH UPDATES
Sales orders, inventory & trends data from multiple online and physical locations
Data is uploaded to S3 for staging
Amazon EMR with spot instances is used to sort, aggregate and join datasets
Processed Data is loaded into Amazon Redshift
Reporting, business apps, and business intelligence
REAL-TIME STREAMING
Real-time data is loaded and processed with Amazon Kinesis Streams
Real-time data is uploaded to DynamoDB
Updated information immediately available online to business users & customers

Data Warehousing



Data Sources

S3

Upload data from multiple sources into S3

EMR

Use Amazon EMR to transform and cleanse the data (ETL)

S3

Load formatted and cleansed data into S3

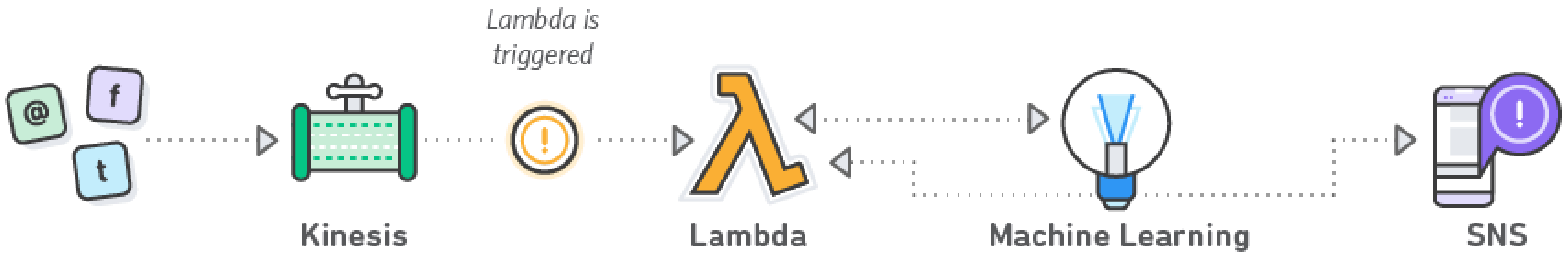
Redshift

Redshift loads data in parallel optimizing it for fast analytics queries

QuickSight

Analyze and visualize data with Amazon Quicksight

Smart Applications | Machine Learning



Lambda is triggered

Kinesis

Lambda

Machine Learning

SNS

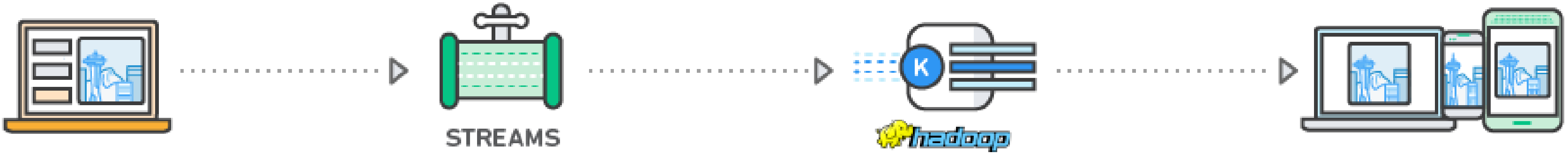
Create an Amazon Kinesis stream for receiving data

Use AWS Lambda to coordinate the data flow

Create an Amazon Machine Learning Model to create real-time predictions

Use Amazon SNS to notify customer support agents

Clickstream Analysis



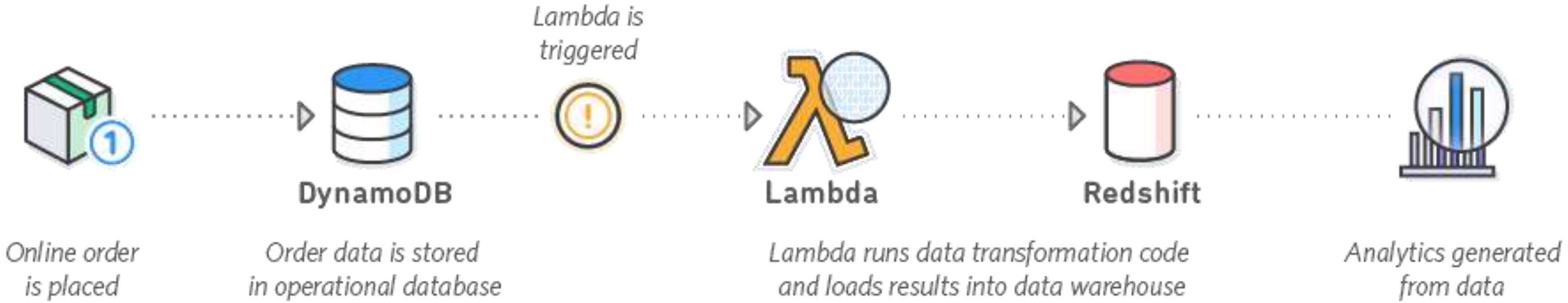
Send clickstream data to Kinesis Streams

Kinesis Streams stores and exposes clickstream data for processing

Custom application built on Kinesis Client Library makes real-time content recommendations

Readers see personalized content suggestions

Event-driven Extract, Transform, Load (ETL)


Lambda is triggered
DynamoDB
Lambda
Redshift
Online order is placed
Order data is stored in operational database
Lambda runs data transformation code and loads results into data warehouse
Analytics generated from data