Architecting an Open Data Lake for the Enterprise
Today’s Presenters

Daniel Geske, Solutions Architect, Amazon Web Services
Armin Wallrab, Director PreSales, Talend
Today’s Agenda

• An overview of AWS and AWS Marketplace, with an emphasis on AWS data lake solutions and Talend
• Overview of the Talend solutions featured in our story
• The Beachbody success story with AWS and Talend
• Q&A/Discussion
Learning Objectives:

1. How to migrate a variety of structured and unstructured data sources to a data lake
2. How to shorten development and testing cycles
3. How to mitigate complex deployment challenges common to real-time data
4. How to take advantage of Spark and Hadoop by generating native code
The Data Lake and AWS

Drive business value with any type of data
Legacy Data Warehouses & RDBMS

- Complex to setup and manage
- Do not scale
- Takes months to add new data sources
- Queries take too long
- Cost $MM upfront
Should I Build a Data Lake?

Starting by **amassing "all your data"** and dumping into a large repository for the data gurus to start finding "insights" is like **trying to win the lottery** by buying all the tickets.
Rethink How to Become a Data-driven Business

• **Business outcomes** - start with the insights and actions you want to drive, then work backwards to a streamlined design

• **Experimentation** - start small, test many ideas, keep the good ones and scale those up, paying only for what you consume

• **Agile and timely** - deploy data processing infrastructure in minutes, not months. take advantage of a rich platform of services to respond quickly to changing business needs
START HERE WITH A BUSINESS CASE
Experiment and Scale Based on Your Business Needs

MATCH
AVAILABLE DATA

Data

Ingest/Collect

Store

Process/analyze

Consume/visualize

Answers & Insights

Metrics and Monitoring
Workflow Logs
ERP Transactions
Business Outcomes on a Modern Data Architecture

**Outcome 1: Modernize and consolidate**
- Insights to enhance business applications and create new digital services

**Outcome 2: Innovate for new revenues**
- Personalization, demand forecasting, risk analysis

**Outcome 3: Real-time engagement**
- Interactive customer experience, event-driven automation, fraud detection

**Outcome 4: Automate for expansive reach**
- Automation of business processes and physical infrastructure
Use an Optimal Combination of Highly Interoperable Services

1. Crawlers scan your data sets and populate the Glue Data Catalog
2. The Glue Data Catalog serves as a central metadata repository
3. Once catalogued in Glue, your data is immediately available for analytics
Why Amazon S3 for Modern Data Architecture?

**Durable**
Designed for 11 9s of durability

**Available**
Designed for 99.99% availability

**High performance**
- Multiple upload
- Range GET

**Easy to use**
- Simple REST API
- AWS SDKs
- Read-after-create consistency
- Event notification
- Lifecycle policies

**Scalable**
- Store as much as you need
- Scale storage and compute independently
- No minimum usage commitments

**Integrated**
- Amazon EMR
- Amazon Redshift
- Amazon DynamoDB
- Amazon Athena
Decouple Storage and Compute

- Legacy design was large **databases** or **data warehouses** with integrated hardware
- Big Data architectures often benefit from **decoupling** storage and compute
More Data Agility with Talend

Populate, Manage and Govern your Data Lake
Stakes are higher than ever with Big Data

- Revenue from Big Data and analytics applications, tools and services: 50%
- Companies that plan on increasing spending on analytics and making data discovery a more significant part of the architecture: 73%
- Big Data projects that will fail to deliver against expectations: 1/2
Why Data Lake projects fail

- Lack of Expertise
- Poor Architectural Design & Integration
- No DevOps Practices for Scalability & Testing
- Siloed Operating Model
- Poor Data Governance
Foundational elements of operating a Data Lake

- Data Classification
- Data Preparation
- Security
- Data Profiling
- Data Integration
- Data Governance
- Metadata Management
- Self-service Data Ingest
### Solution Architecture of a Data Lake

#### Data Sources
- **Structured**
  - ERP
  - Relational
  - Mainframe
- **Unstructured**
  - Machine sensor data
  - Web logs
  - Documents
- **External Social**
  - Sentiment
  - Vigilance
  - Trend shift

#### Data Ingestion
- **Batch**
- **Real Time**

#### Data Lake
- **Raw Layer**
  - Source data ready for consumption by Data Scientists, Data Stewards and ETL Developers
- **Processing Layer**
  - Standardize on Corporate Governance
  - Apply Data Quality Rules and Policies
- **Refined Layer**
  - Apply Business Rules
  - Transformed Data required by Lines of Business Specific Views
  - Conformed data objects for business data repositories

#### Data Governance
- Metadata Management

#### Business Data Repositories
- Regional Data Marts

#### Advanced Analytics & Reporting
- Qlik

---

**[Diagram Image]**
Graphical Integration Design with Talend Studio

- ~1,000 visual drag-and-drop components
  - Databases, Cloud, Big Data, Files, Applications, Transformations, Processing, Machine Learning, Data Quality, ...
- Native code generation for Java & Spark
  - Run anywhere
- Standard Software Development Lifecycle
  - Continuous Integration
- (Big) Data Quality included
  - Profiling, Standardization, Address Correction, De-duplication, Masking, ...
- Self-Service for Line of Business
  - Data Preparation & Data Stewardship
AWS support in Talend

Storage
- Amazon S3
- Amazon DynamoDB

Database
- Amazon RDS
- Amazon Redshift

Computation
- Amazon EMR
- Amazon EC2

+ Cluster Management

Messaging
- Amazon Kinesis
- Amazon SQS

Notification
- Amazon SNS
- Amazon SES

Others
- Amazon IAM
- Amazon QuickSight
Talend Data Prep: Self-Service for the Business Analyst
Graphical Integration Design with Talend Studio
Flexible deployment with Talend

On-Premises

- Talend Data Integration
- Talend Big Data
- Firewall

Multi-Tenant Cloud iPaaS

- Talend Cloud

Public Cloud

AWS

Talend Studio

Develop in Talend Studio – Deploy anywhere
A unified Platform for integrating the Data Lake
Beachbody – Fitness goes Big Data

Driving innovation with Talend on AWS
About Beachbody

• A leading provider of **fitness, nutrition, and weight-loss programs**
• Operates with **800+ employees**
• Empowers over **23 Million customers**

"We can understand in near-real-time consumer behavior in fitness centers with a cloud big data lake and self-service analytics. The benefits are increased effectiveness for our digital marketing campaigns, and decreased customer churn."

*Eric Anderson, Executive Director, Data, Beachbody LLC*
The Challenge - Do More Better, Faster, Cheaper

- Revenue
- Contribution Margin
- Returns
- Customer Count
- EBITDA
- Sales
- Average Subscription Cycles
- Streaming Customers

WHAT WE KNOW...

The Rest...

Need a repository to capture and store the massive amount of data we’re currently ignoring.

How can we personalize offers for each customer?

How can we better target and retarget customers leveraging customer purchase and behavioral data?

How can I better match customers with coaches to ensure longevity and mutual success?

What indicators determine if a subscription is about to cancel?
The Technology
Data Lake Component - Storage

- Amazon S3
- All data is encrypted at rest
- Five categories:
  - **Raw**
  - **Pre-Processed** - for efficient consumption
  - **Processed** - Curated data with business rules applied
  - **Sensitive** - Encrypted zone
  - **Work** – Sandbox for projects
Data Lake Component – Data Pipeline

- Executes processes natively in Talend with Push-Down
- Leverages cluster security
- Will be used to:
  - Ingest & land raw data
  - Transform raw data
  - Orchestrate workflow for data science & analyses
  - Load data into RDBMS

Vendors were compared by capabilities across 6 categories
Data Lake Component – RDBMS

- Amazon Redshift or Presto for analytical database usage?
- Hive for ad-hoc queries against fine-grained data
- Data from “processed” storage is loaded into analytical DB
- Business users connect to RDBMS with data viz or query tools
- Will serve as Enterprise Data Warehouse platform
- All data encrypted at rest and in transit
Data Lake Component – Compute

- Hortonworks HDP used for persistent cluster
  - Available 24/7 to business user community
  - Recent versions of Hadoop ecosystem components
  - Open-source platform tracking closely to Apache releases
- Amazon EMR used for transient clusters
  - Experimental and isolated workloads
  - Optimized for rapid cluster creation and tear-down

Vendors were compared by capabilities across 6 categories
Data Lake Component – Analytics

- Spark compute engine with Machine Learning libraries
- Data science models are run in Spark engine
- Python and Scala used for programming
- Results are landed back into storage for further action
- Data pipeline leverages Spark engine for transformations
- Spark engine encrypts data at rest and in motion
Business Benefits

- Reduced Data Acquisition Time by 5x
- Improved Marketing Campaigns
- Reduced Site Tagging Costs
- Improved Employee Retention and Satisfaction
- Automated Customer Self-Service Order Status
- Identified Web Funnel Conversion Opportunities (testing now)
Next steps and further information

• Data Lake on AWS Quick Start:

• Take a Free 30-Day Trial of Talend Cloud:
  https://iam.eu.integrationcloud.talend.com/idp/trial-registration