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

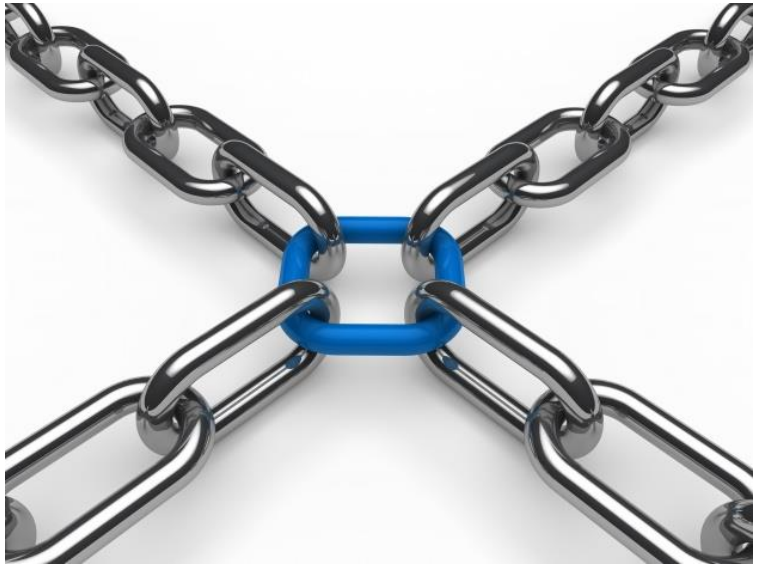
Learning Objectives:

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

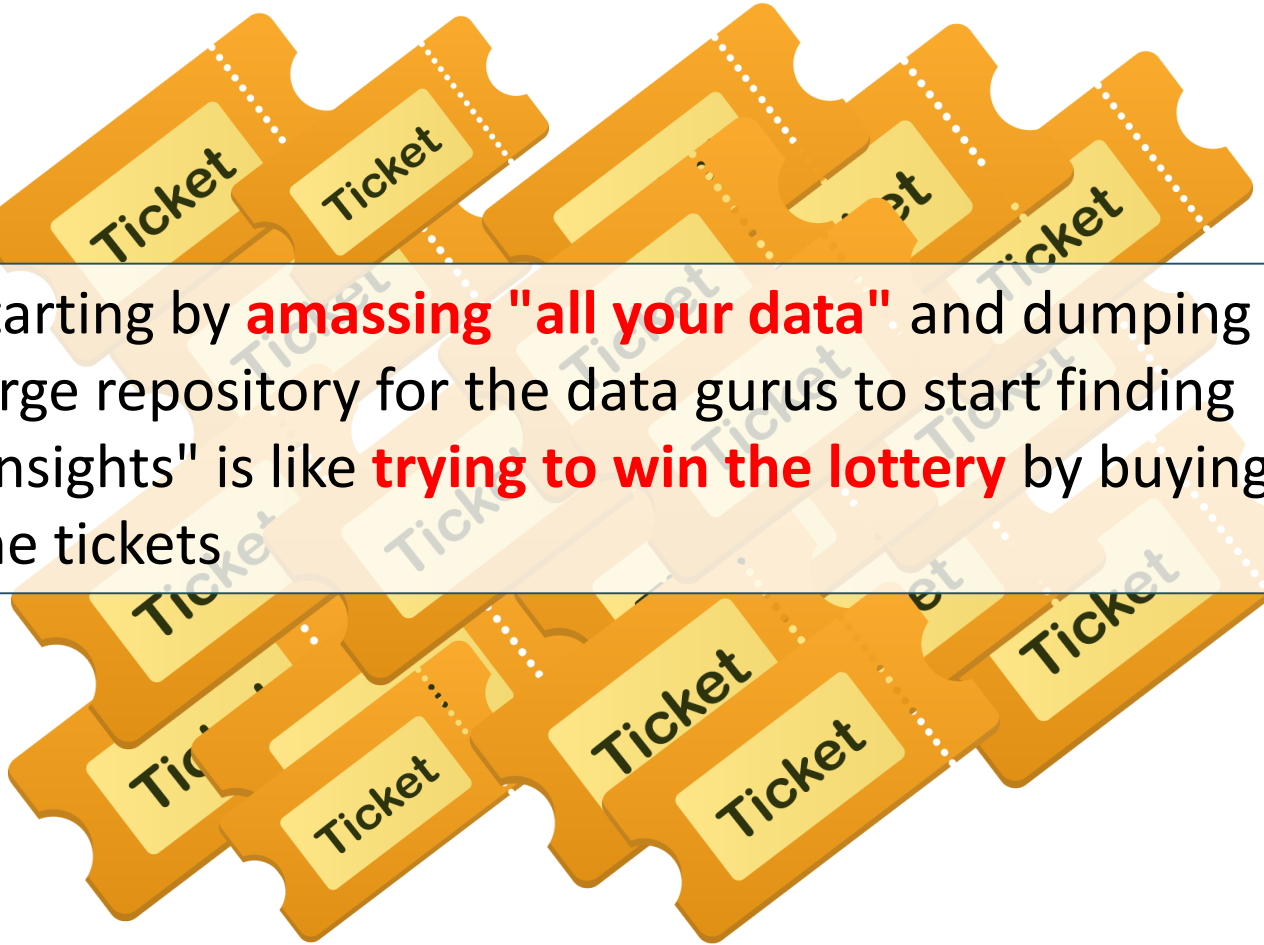
Legacy Data Warehouses & RDBMS



- **Complex** to setup and manage
- Do not **scale**
- **Takes months** to add new data sources
- Queries take **too long**
- **Cost** $MM upfront

aws

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

aws

Building a Data Lake on AWS

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

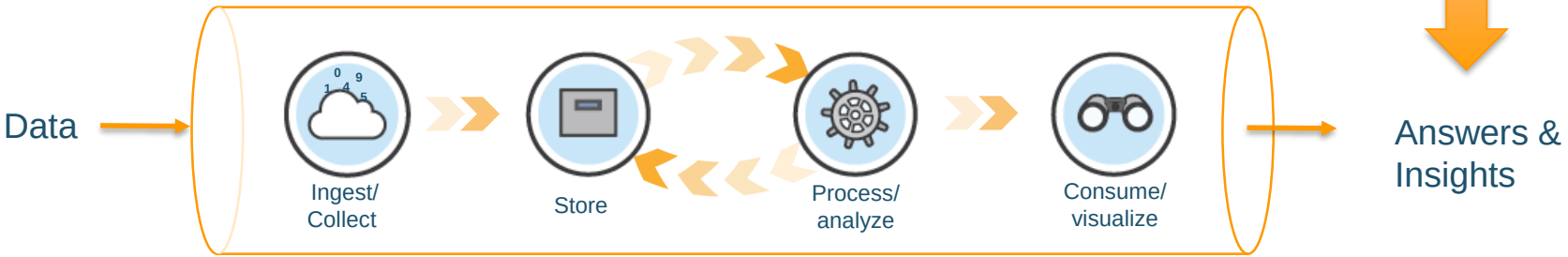
- **Experimentation** - start small, test many ideas, keep the good ones and scale those up, paying only for what you consume

- **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

Business Case Determines Platform Design

START HERE

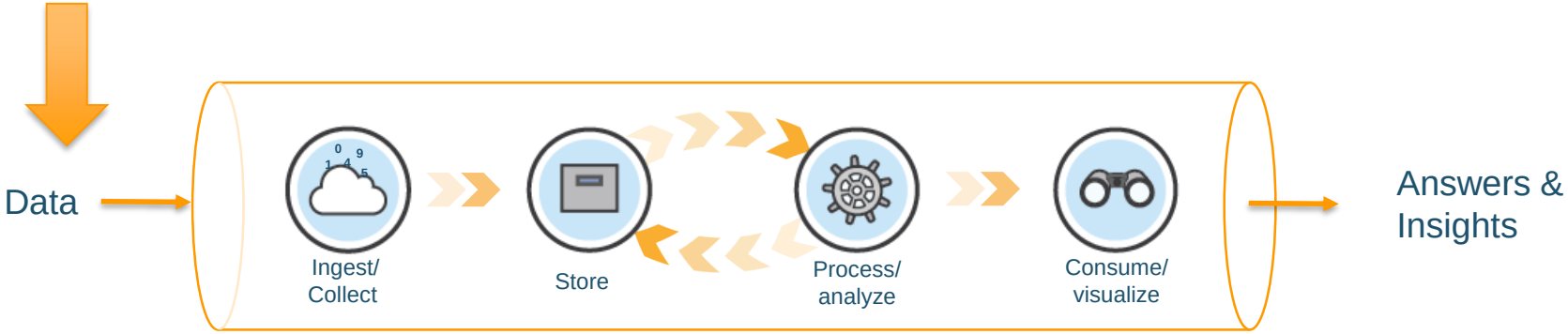
WITH A BUSINESS CASE



Experiment and Scale Based on Your Business Needs

MATCH

AVAILABLE DATA



Metrics and Monitoring

Workflow Logs

ERP Transactions

aws

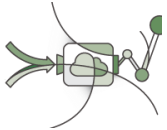
Business Outcomes on a Modern Data Architecture

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

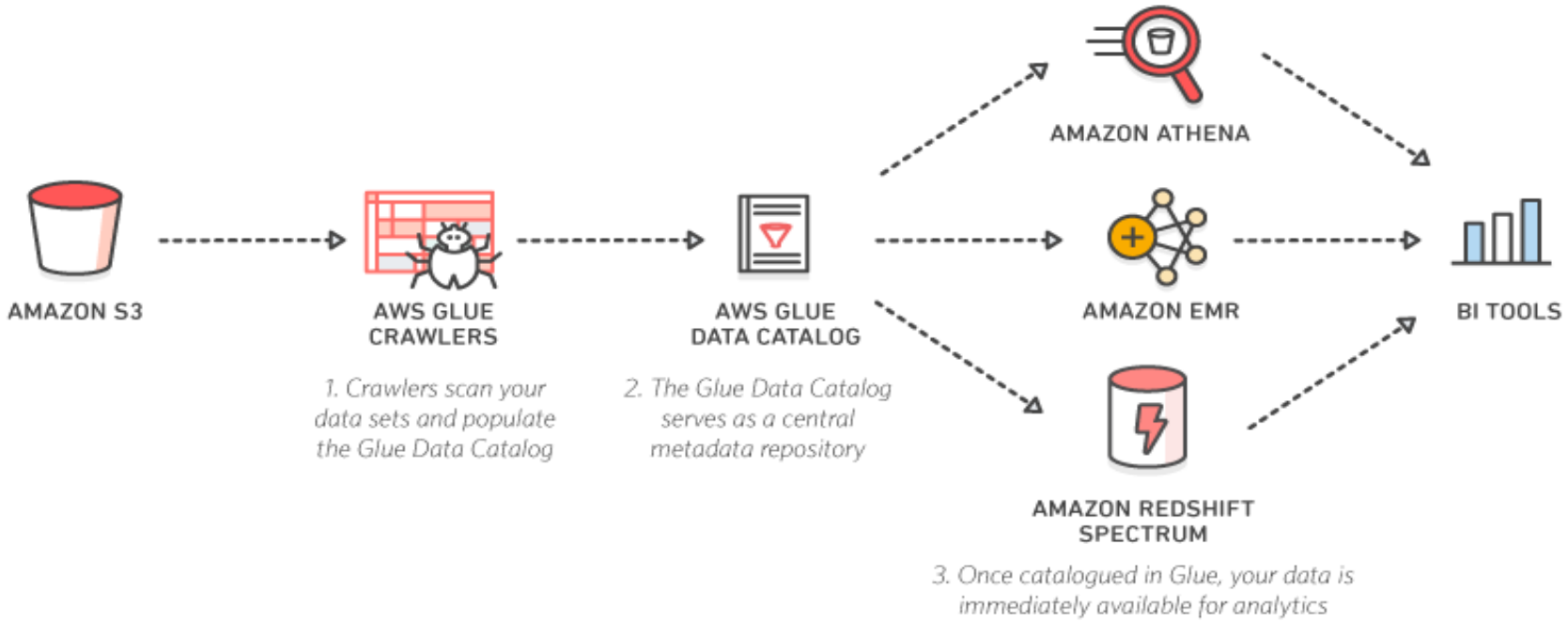
- Interactive customer experience, event-driven automation, fraud detection

Outcome 4 : Automate for expansive reach

- Automation of business processes and physical infrastructure

aws

Use an Optimal Combination of Highly Interoperable Services



Why Amazon S3 for Modern Data Architecture?

Durable

Designed for 11 9s of durability

Available

Designed for **99.99**% availability

High performance

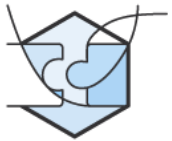
- Multiple upload
- Range GET

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



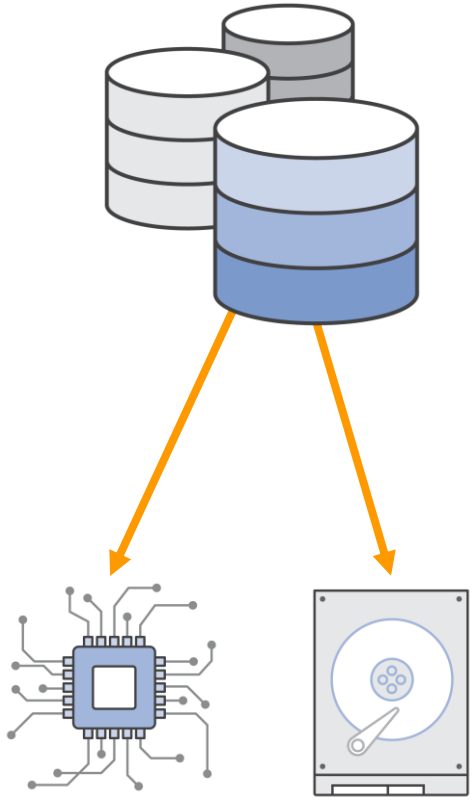
Integrated

- Amazon EMR
- Amazon Redshift
- Amazon DynamoDB
- Amazon Athena

aws

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

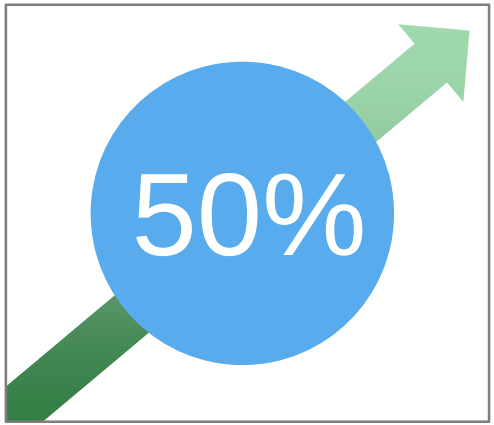


aws

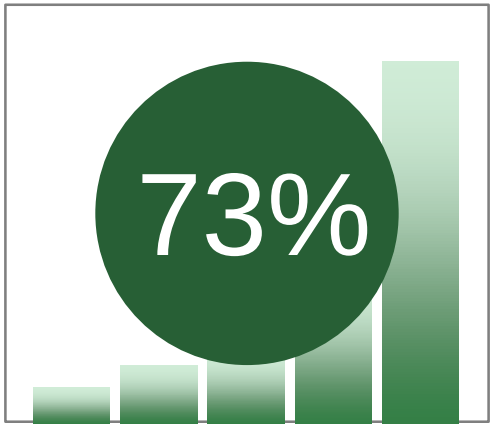
More Data Agility with Talend

Populate, Manage and Govern your Data Lake

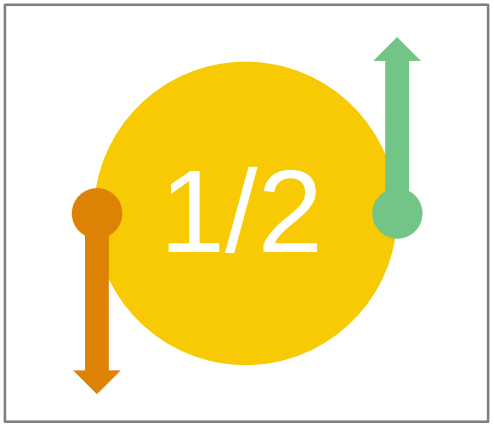
Stakes are higher than ever with Big Data



Revenue from Big Data and analytics applications, tools and services



Companies that plan on increasing spending on analytics and making data discovery a more significant part of the architecture



Big Data projects that will fail to deliver against expectations

talend

aws

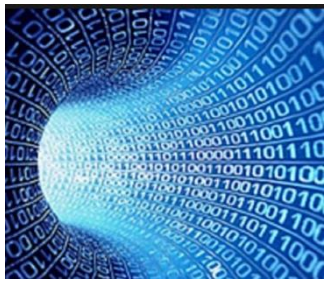
Why Data Lake projects fail

Lack of Expertise

Poor Architectural Design & Integration

No DevOps Practices for Scalability & Testing

Siloed Operating Model

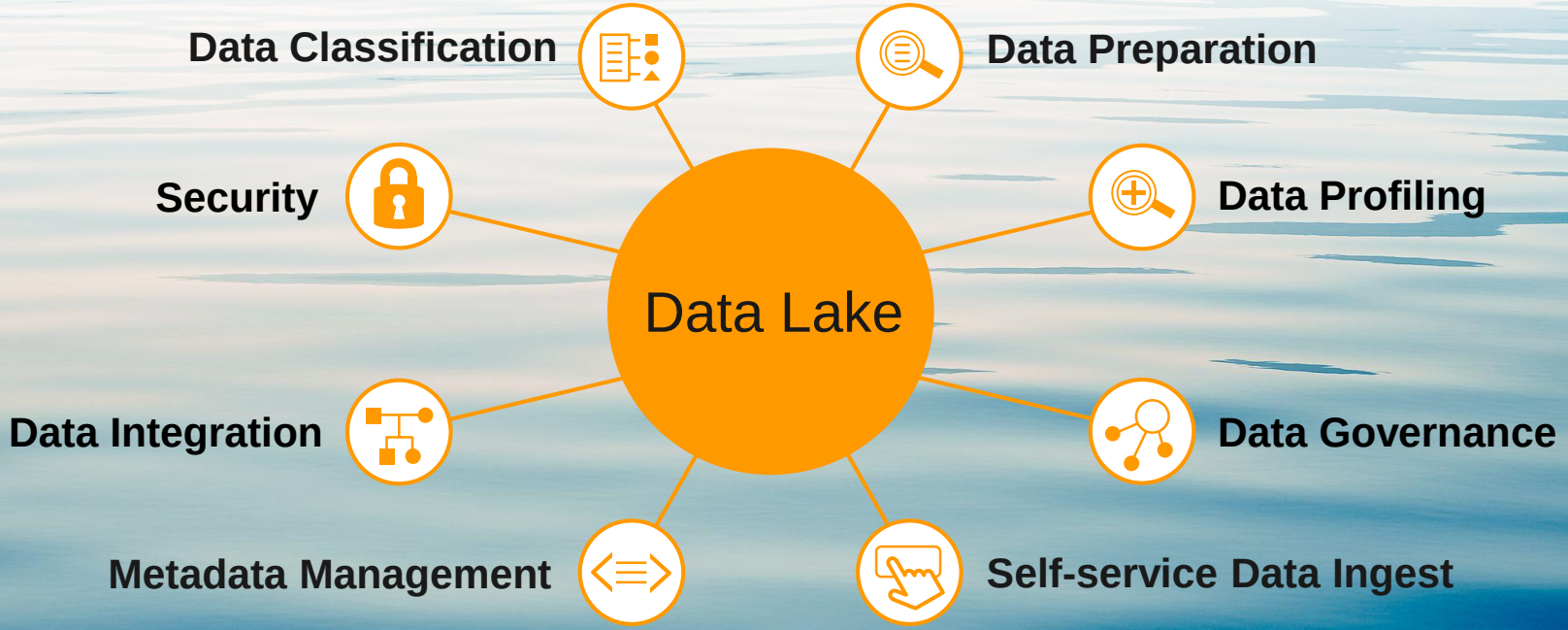


Poor Data Governance

talend

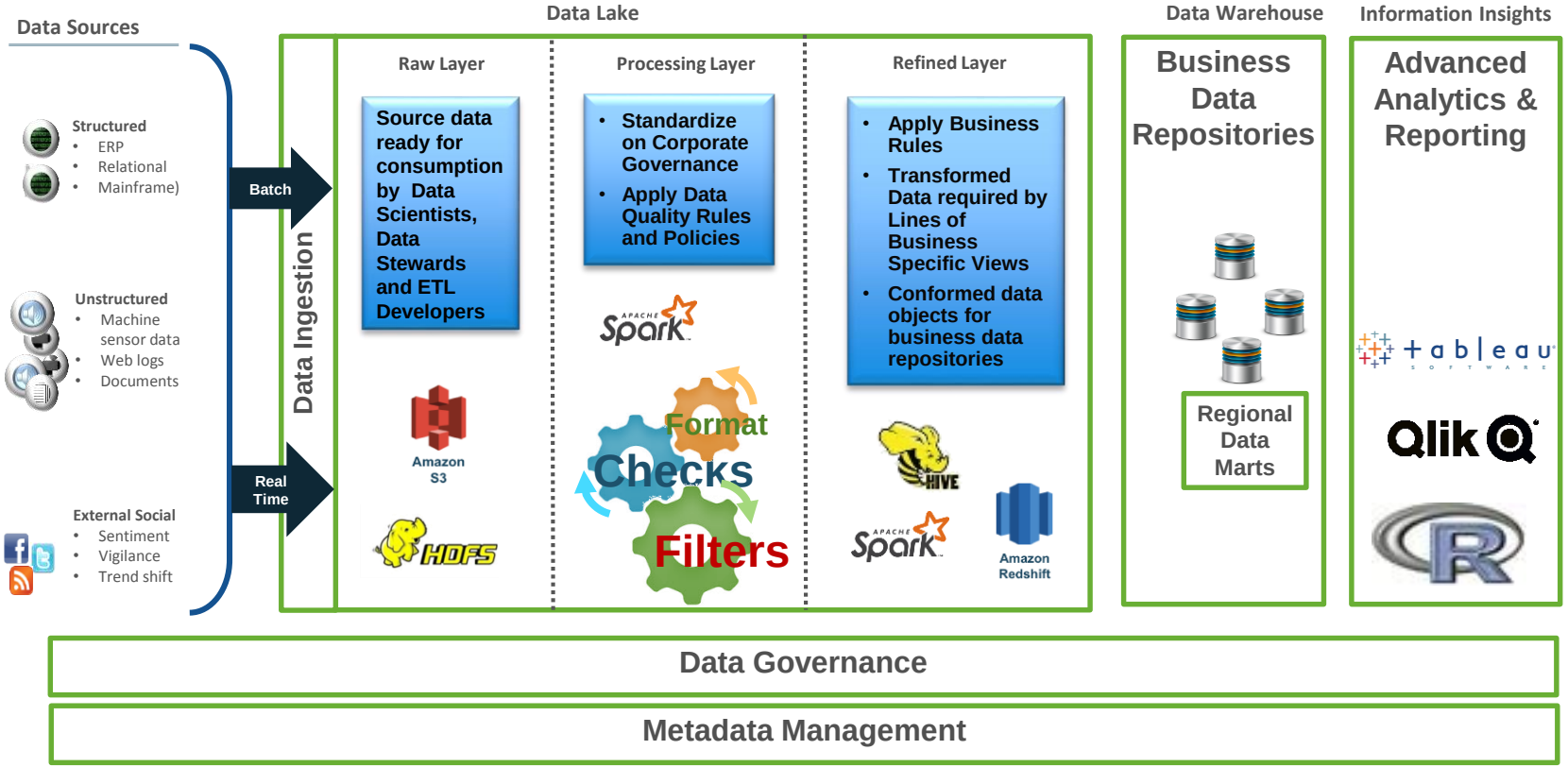
aws

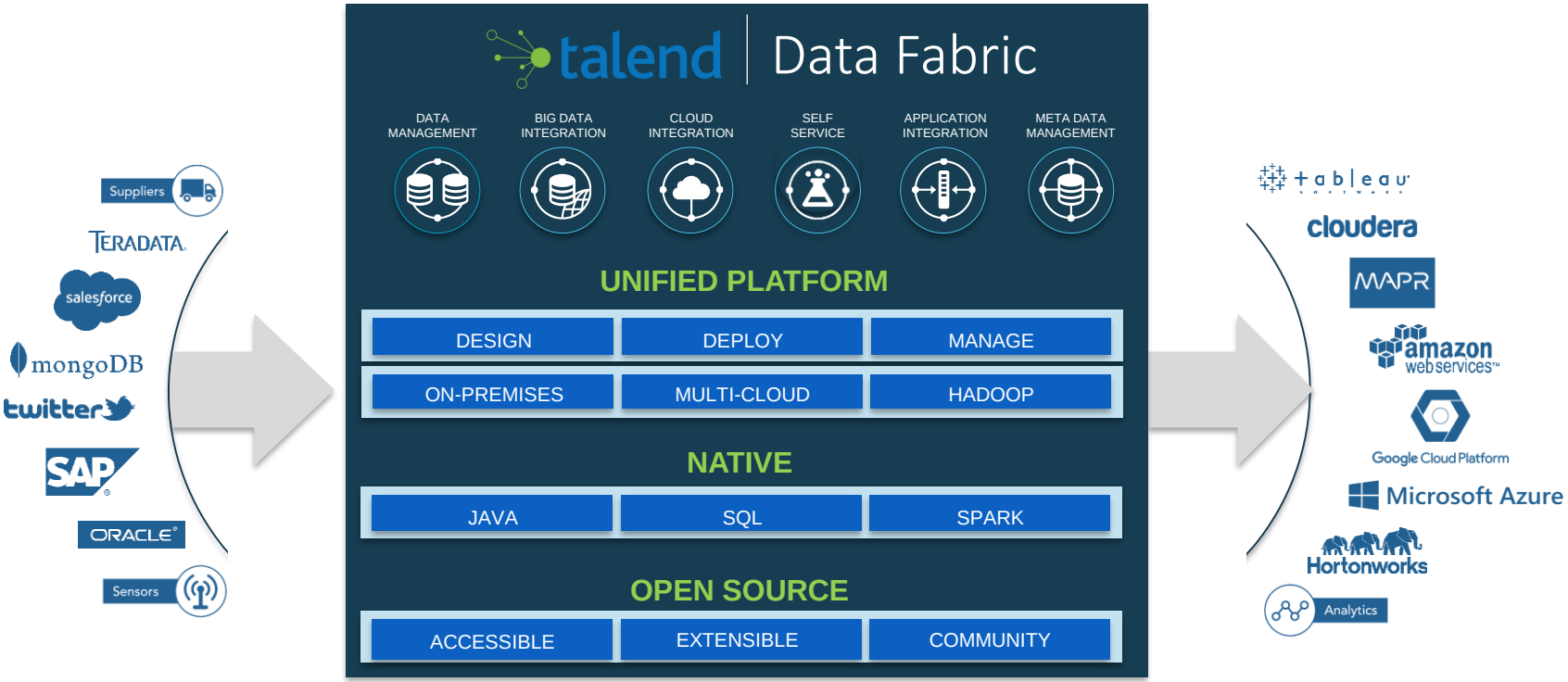
Foundational elements of operating a Data Lake



talend

Solution Architecture of a Data Lake

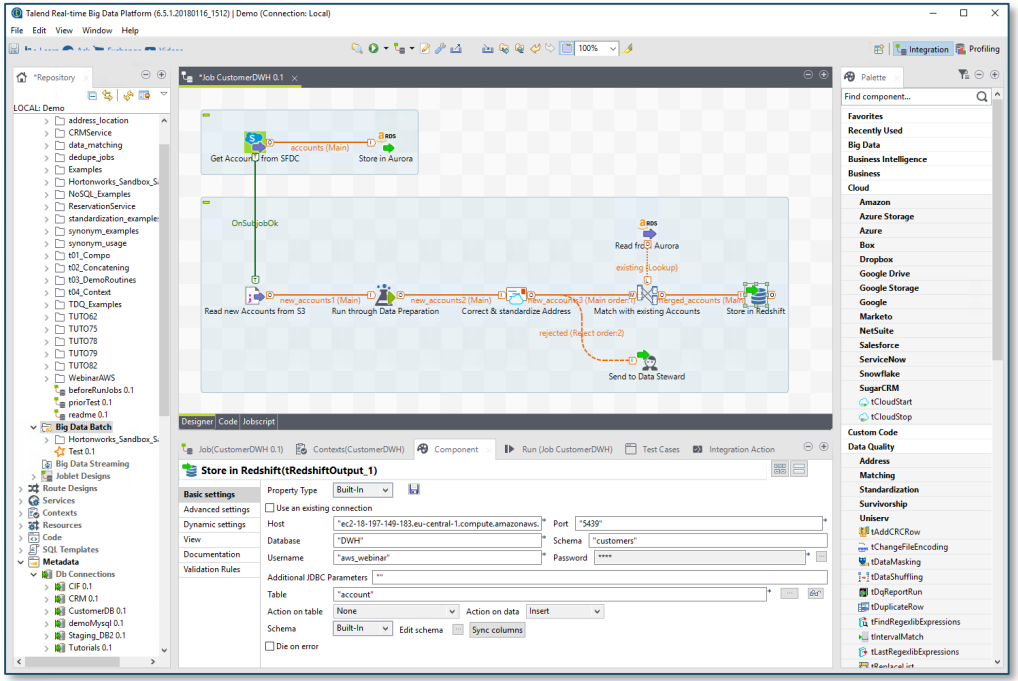



talend
Data Fabric
DATA MANAGEMENT
BIG DATA INTEGRATION
CLOUD INTEGRATION
SELF SERVICE
APPLICATION INTEGRATION
META DATA MANAGEMENT
UNIFIED PLATFORM
DESIGN
DEPLOY
MANAGE
ON-PREMISES
MULTI-CLOUD
HADOOP
NATIVE
JAVA
SQL
SPARK
OPEN SOURCE
ACCESSIBLE
EXTENSIBLE
COMMUNITY
Suppliers
TERADATA
salesforce
mongoDB
twitter
SAP
ORACLE
Sensors
tableau
cloudera
MAPR
amazon web services
Google Cloud Platform
Microsoft Azure
Hortonworks
Analytics

talend

aws

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

Spark hadoop Java talend aws

AWS support in Talend

Advanced Technology Partner

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

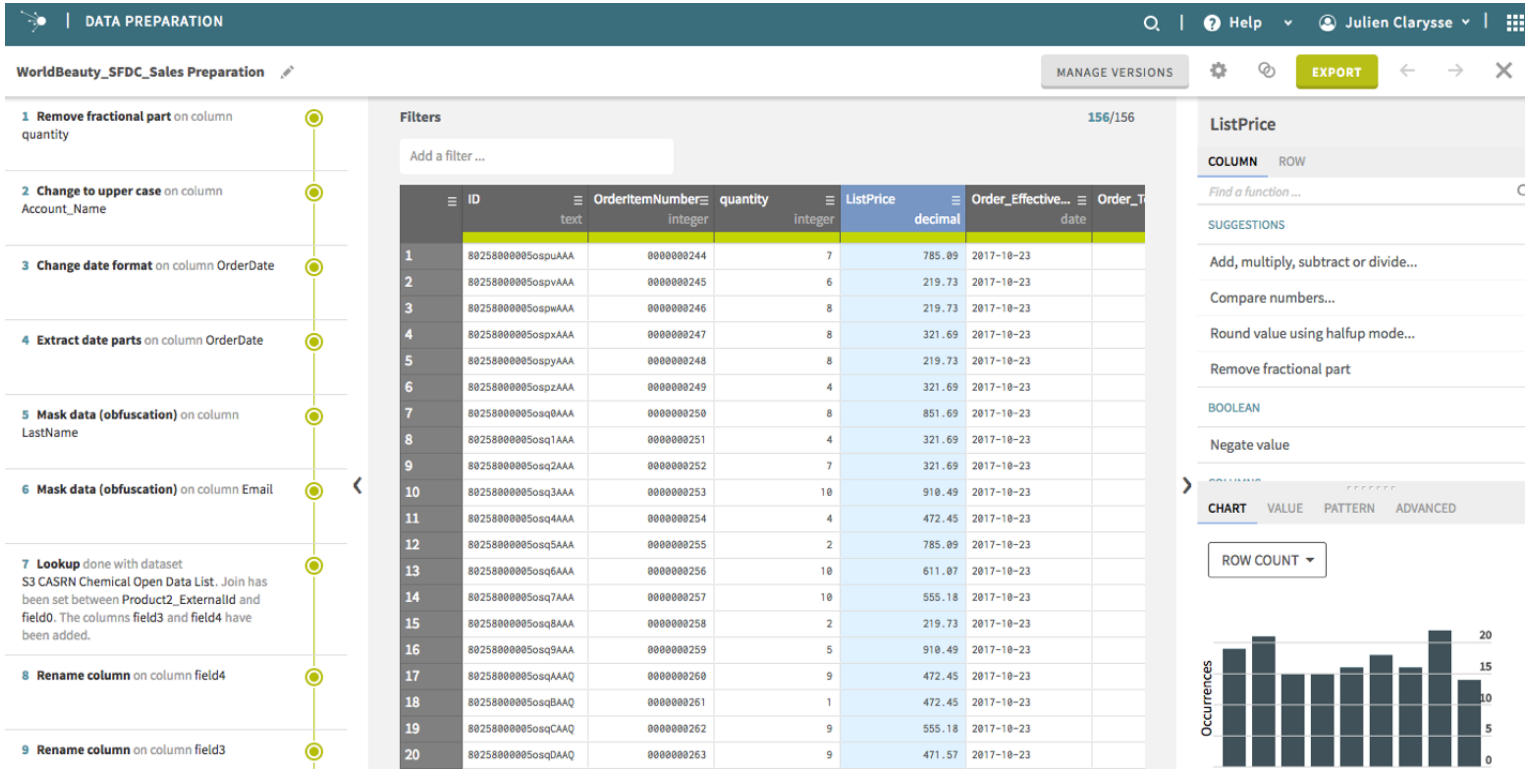

Amazon IAM


Amazon QuickSight

talend

aws

Talend Data Prep: Self-Service for the Business Analyst

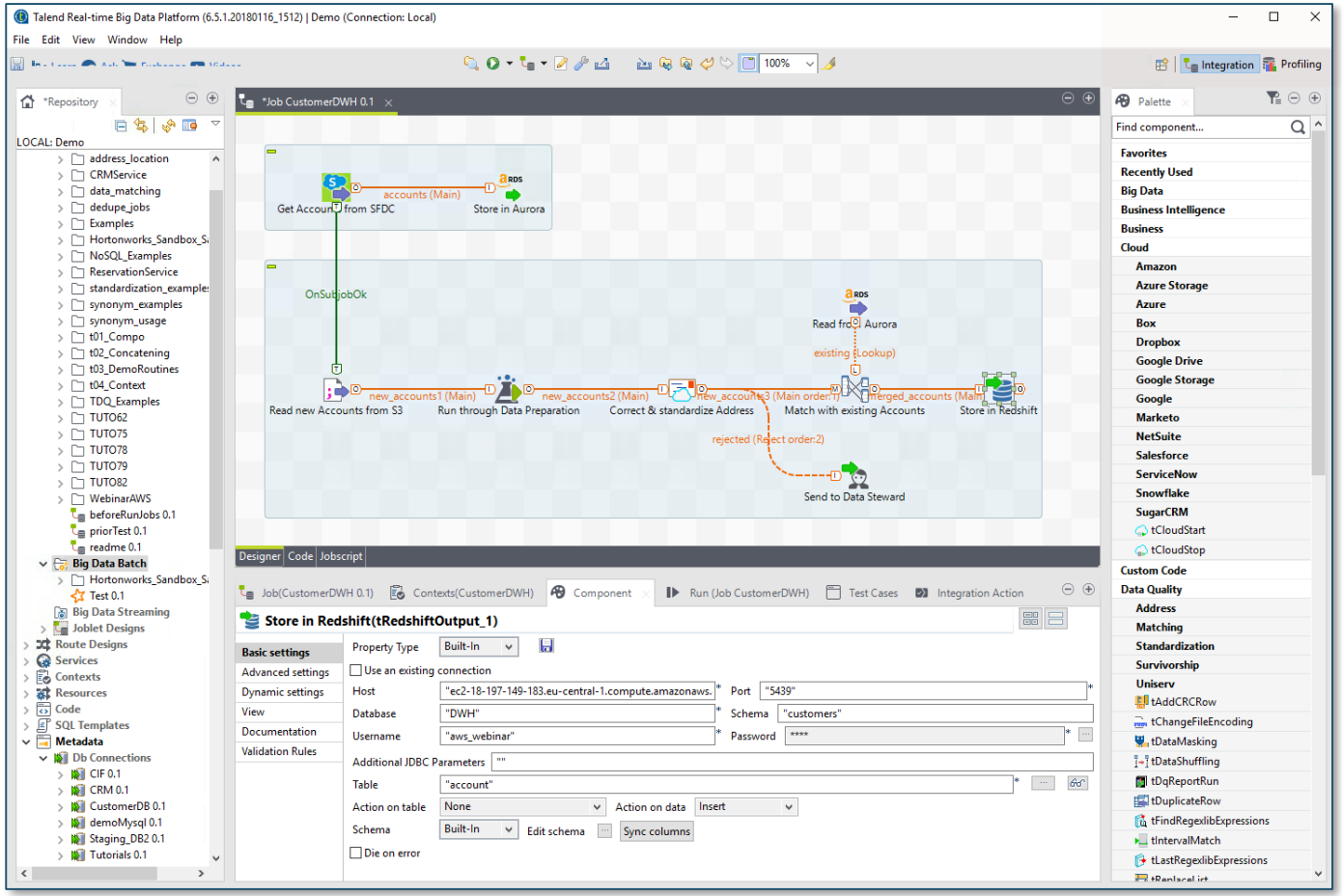


talend

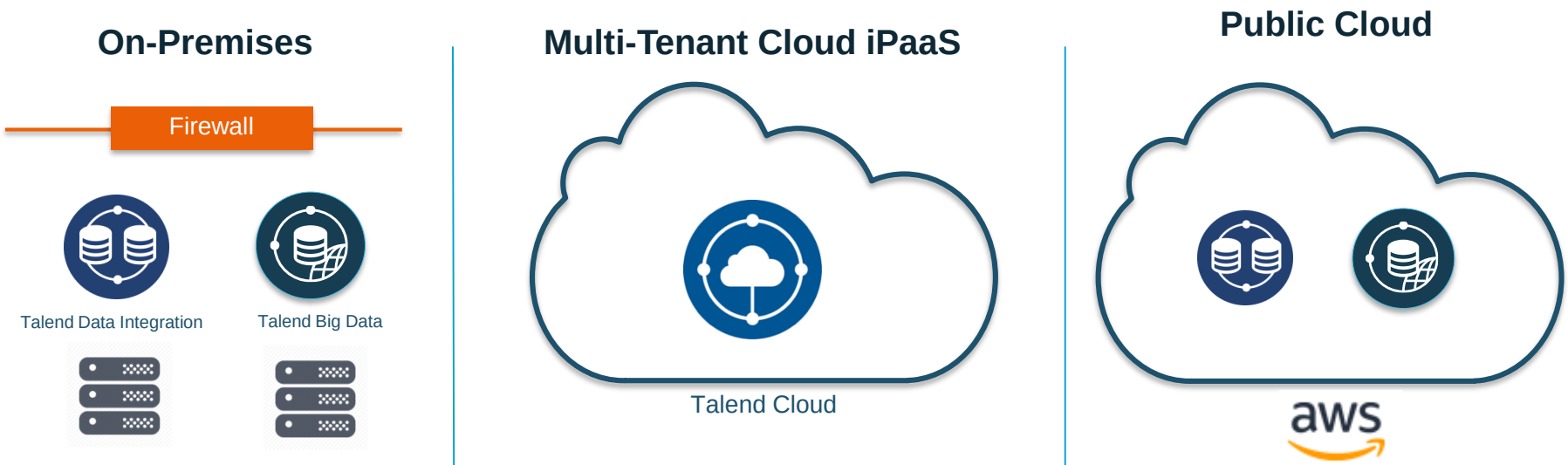
CSV AWS S3 Spark

aws

Graphical Integration Design with Talend Studio



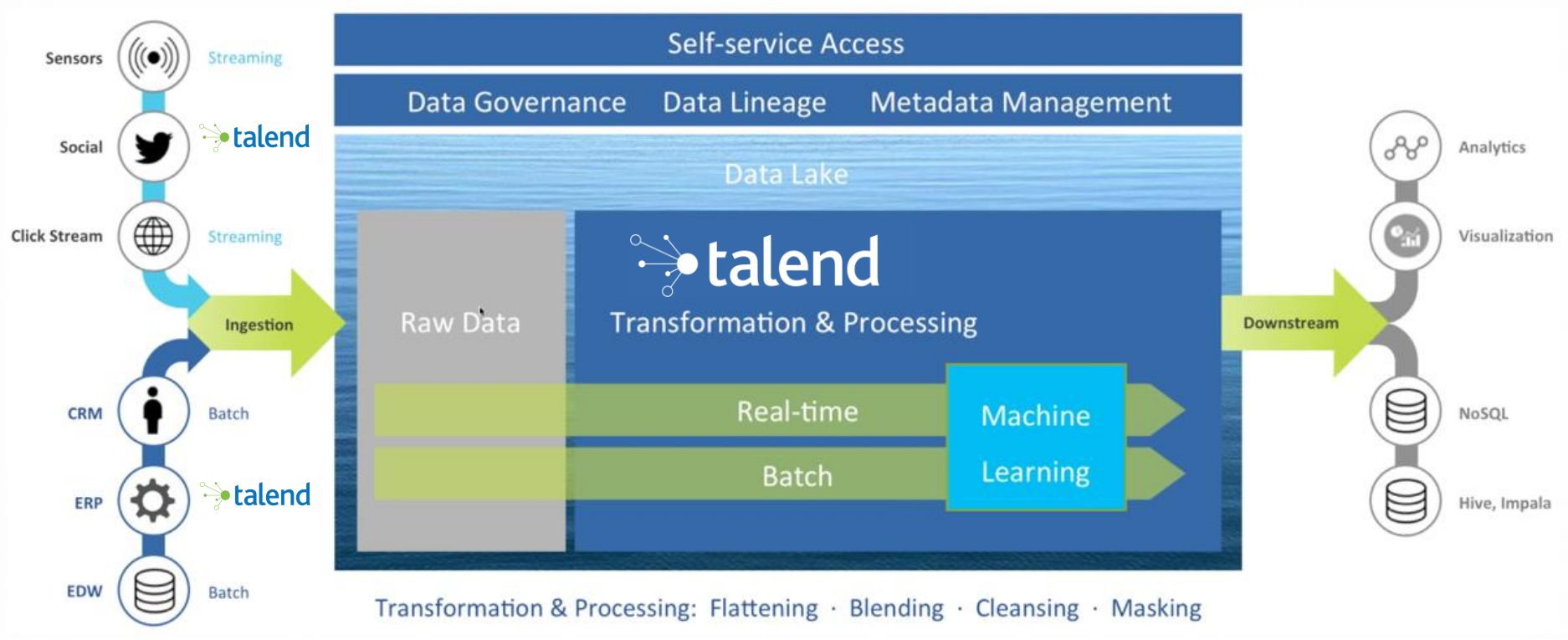
Flexible deployment with Talend



talend

aws

A unified Platform for integrating the Data Lake



talend

aws

Beachbody – Fitness goes Big Data

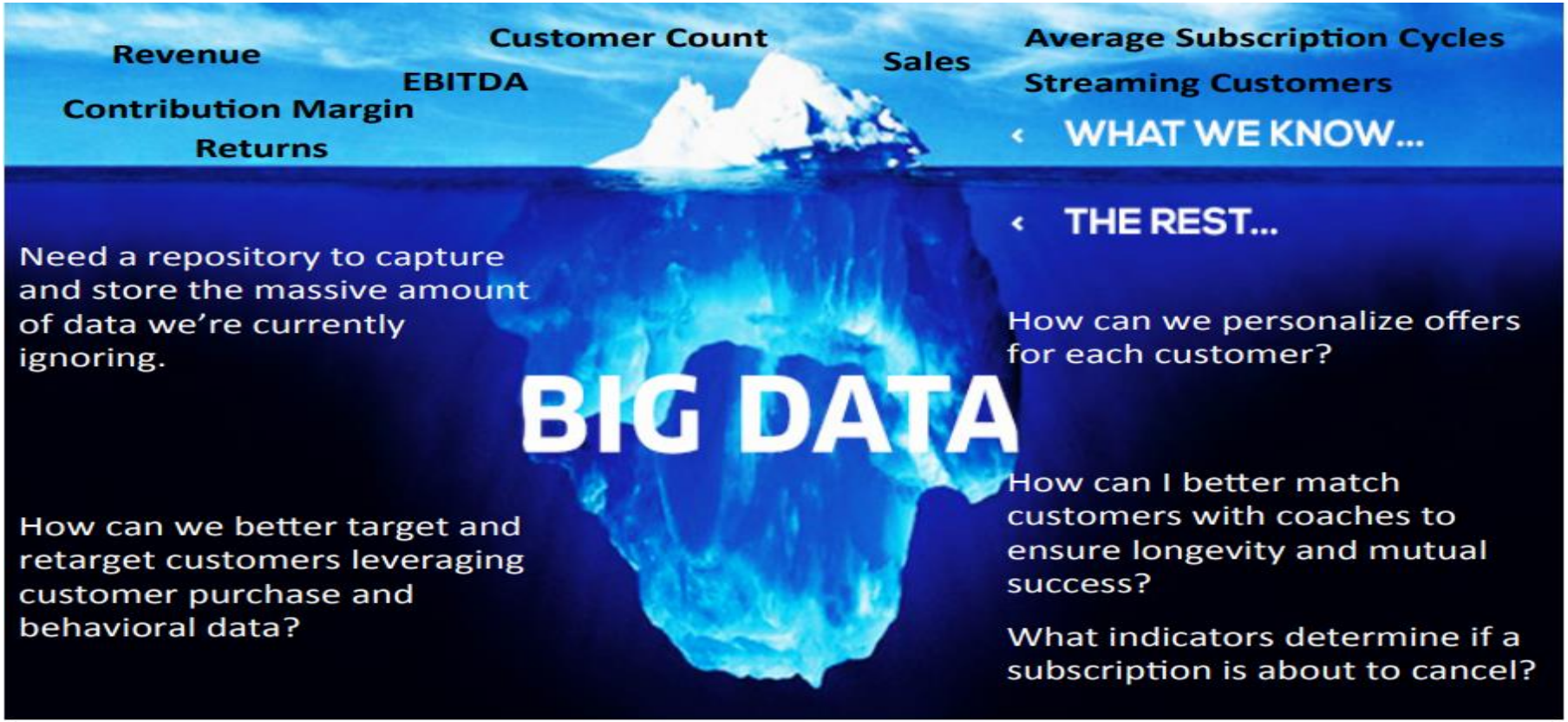
Driving innovation with Talend on AWS

About Beachbody

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

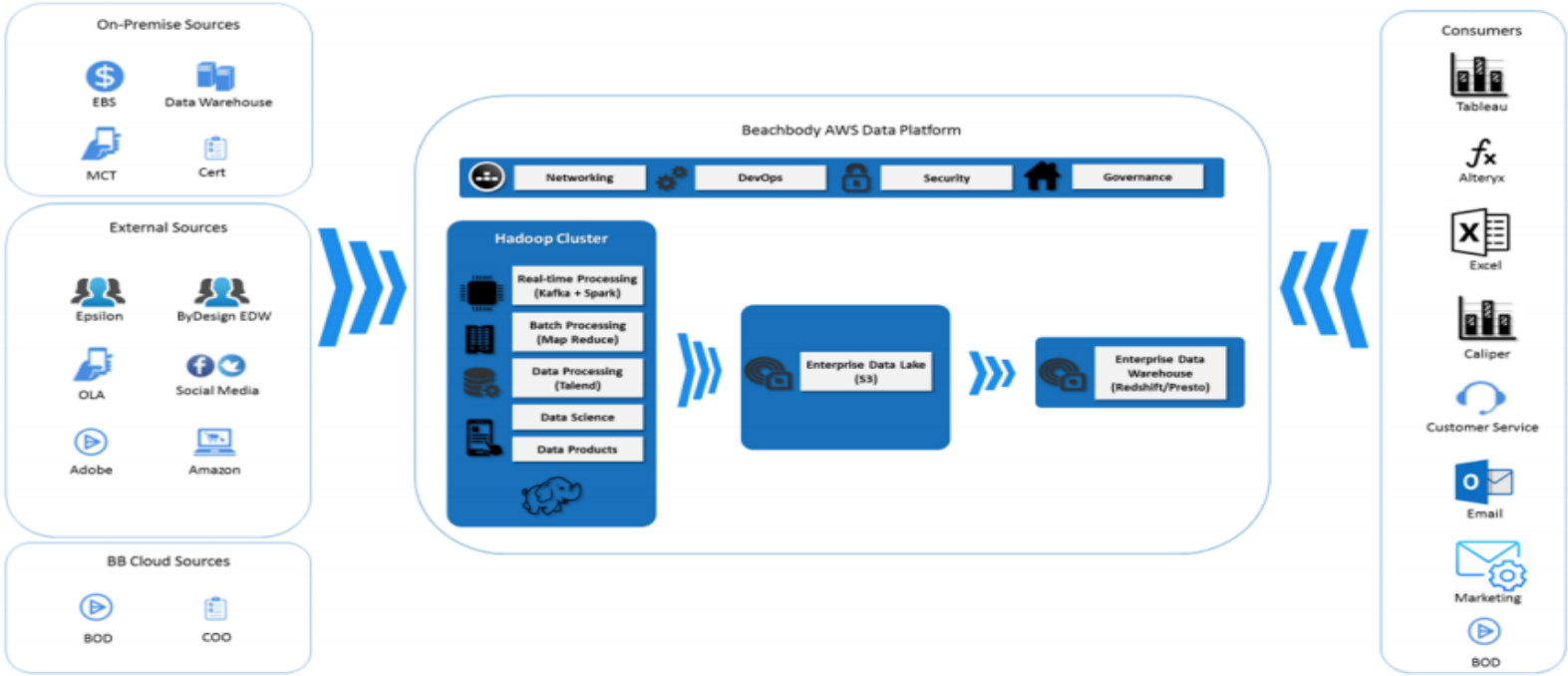
talend BEACHBODY aws

The Challenge - Do More Better, Faster, Cheaper



talend BEACHBODY aws

The Technology



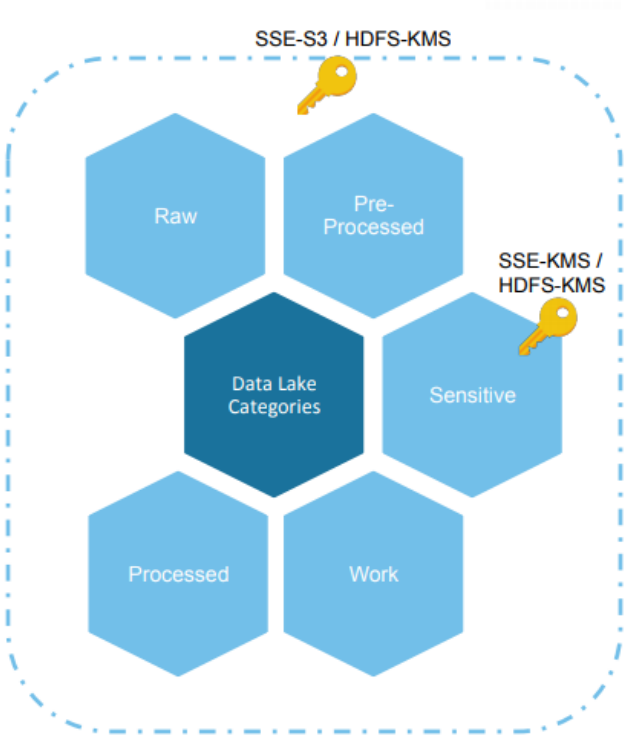
talend

BEACHBODY®

aws

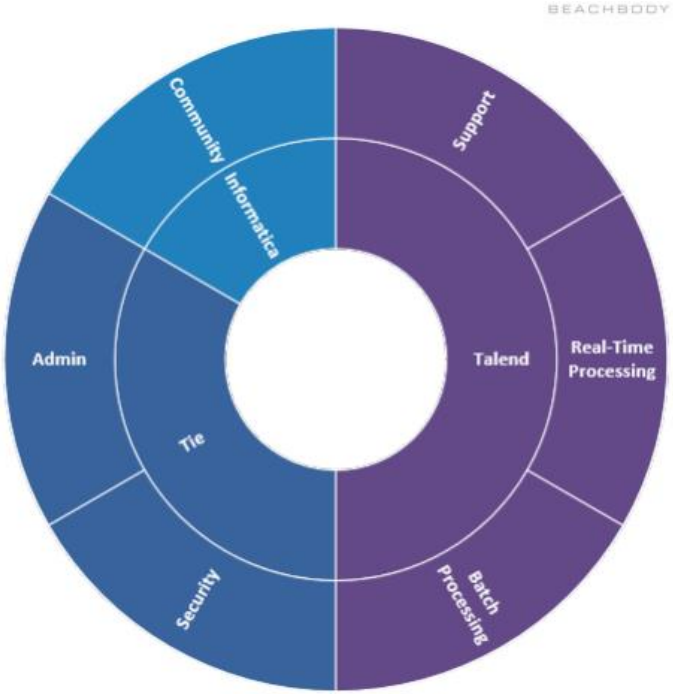
Data Lake Component - Storage

- 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

Amazon **Redshift**

talend

BEACHBODY

aws

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: https://aws.amazon.com/quickstart/architecture/data-lake-with-talend-big-data-platform/

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