



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

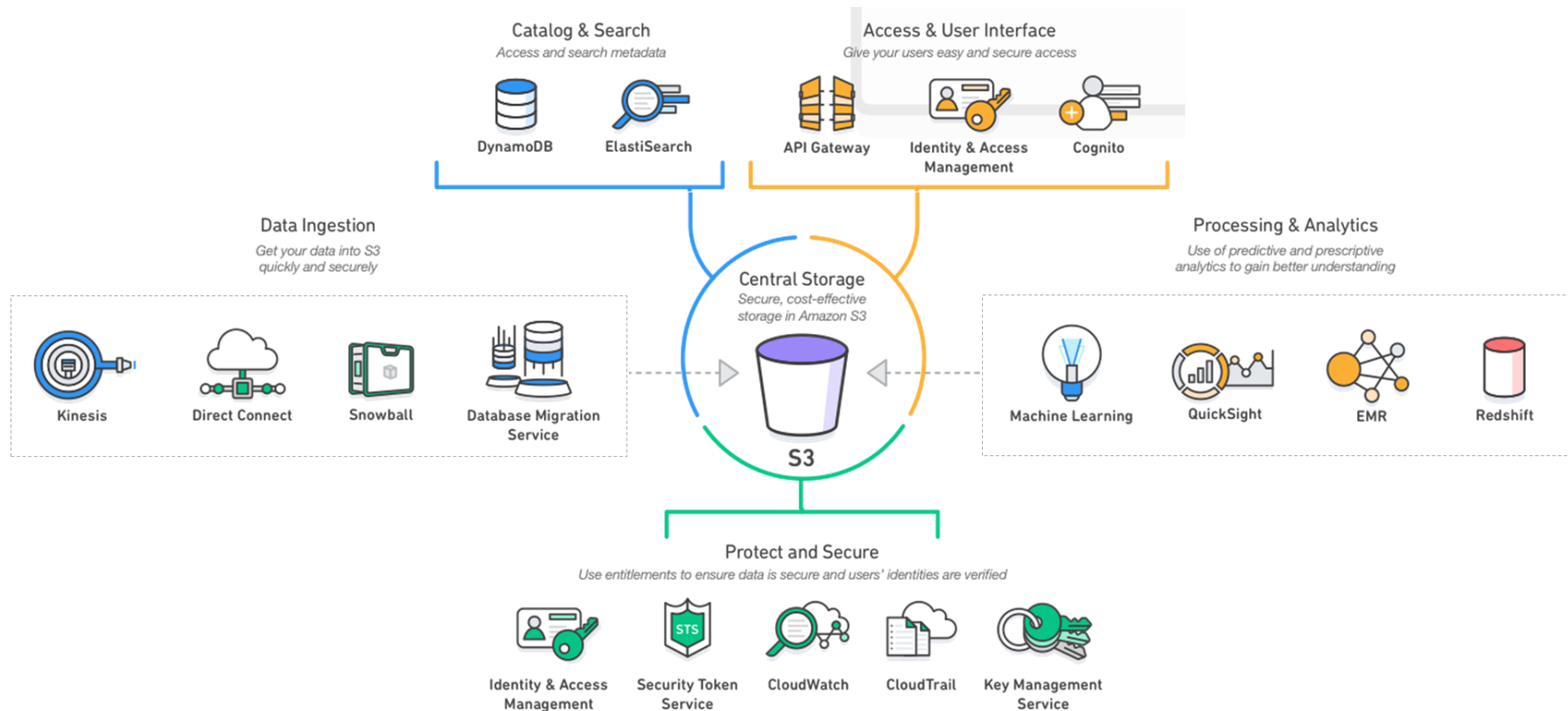


- **Complex** to setup and manage
- Do not **scale**
- **Takes months** to add new data sources
- Queries take **too long**
- **Cost \$MM** upfront

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

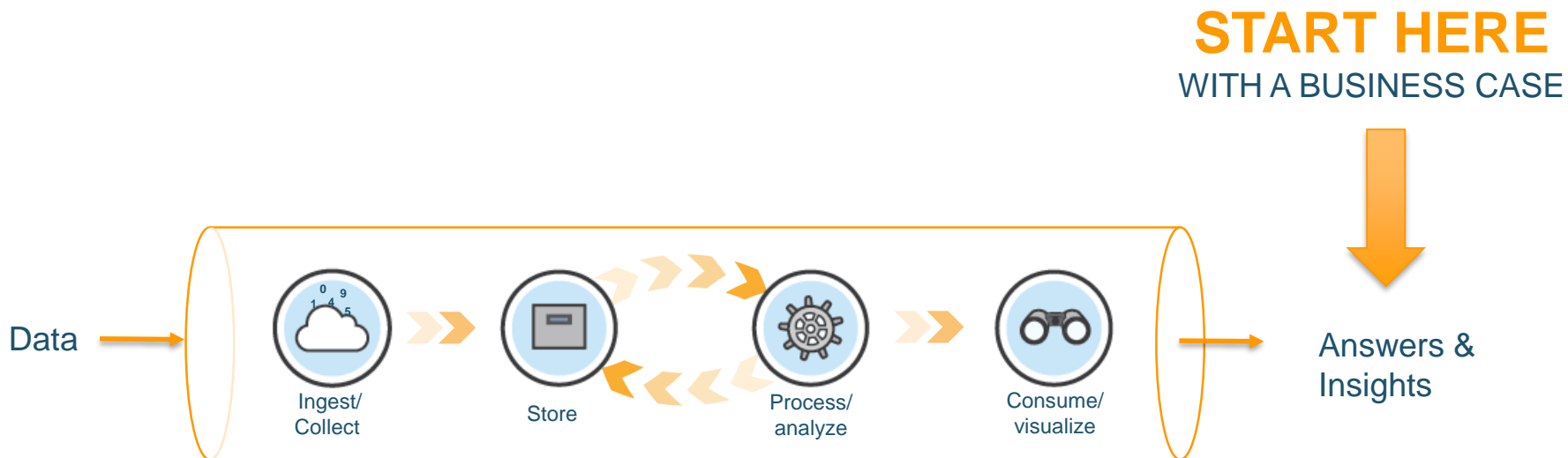
Building a Data Lake on AWS



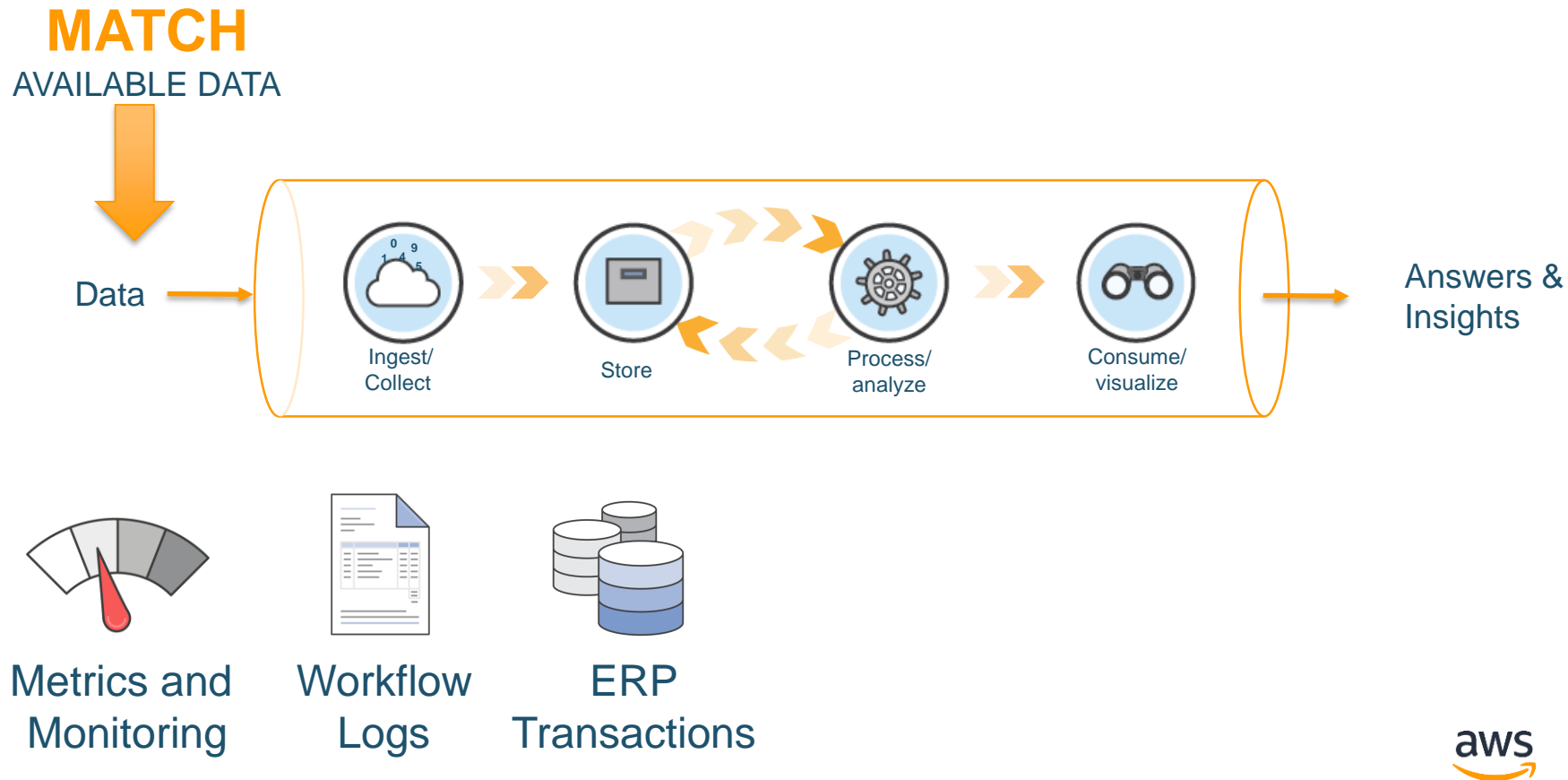
Rethink How to Become a Data-driven Business

- **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



Experiment and Scale Based on Your Business Needs

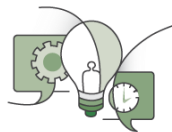


Business Outcomes on a Modern Data Architecture



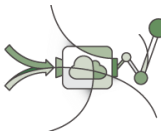
Outcome 1 : Modernize and consolidate

- 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

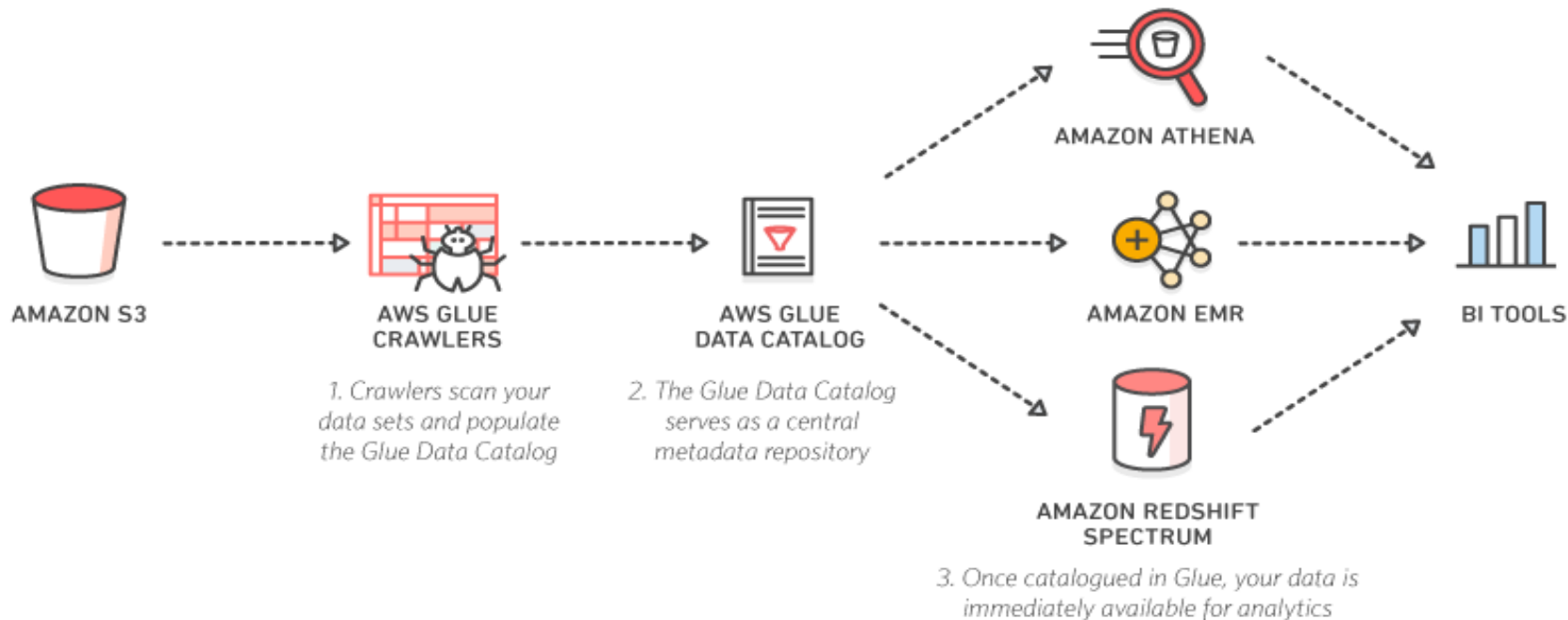
- Interactive customer experience, event-driven automation, fraud detection



Outcome 4 : Automate for expansive reach

- Automation of business processes and physical infrastructure

Use an Optimal Combination of Highly Interoperable Services



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

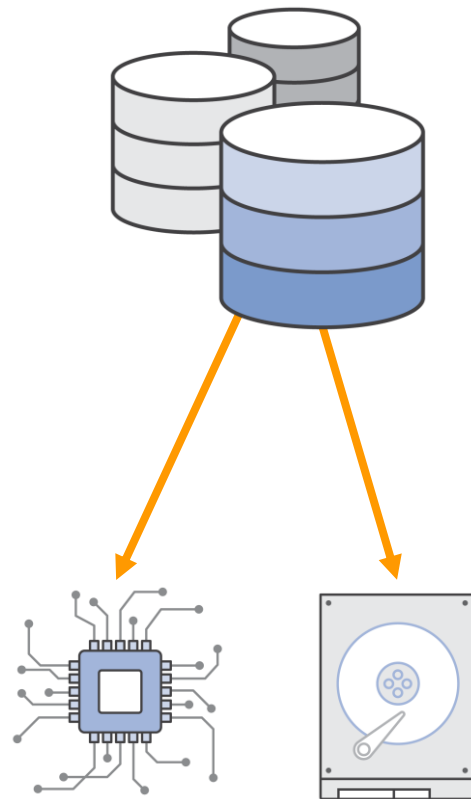


Integrated

- 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

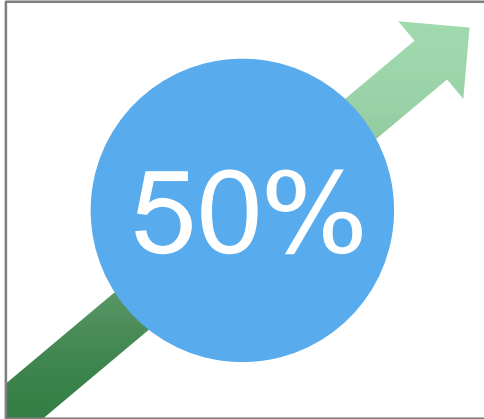




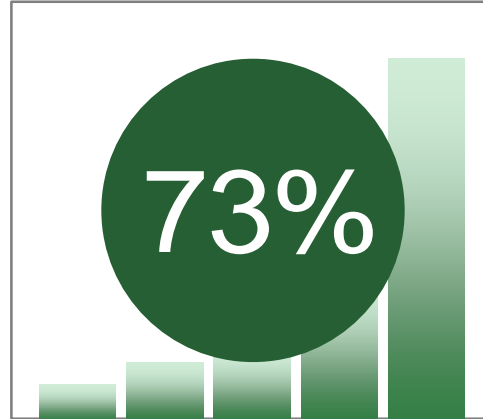
More Data Agility with Talend

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



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

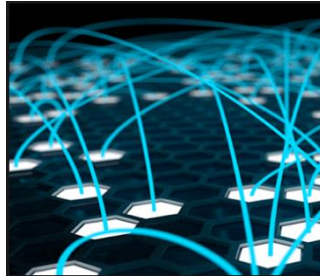
Why Data Lake projects fail



Lack of Expertise



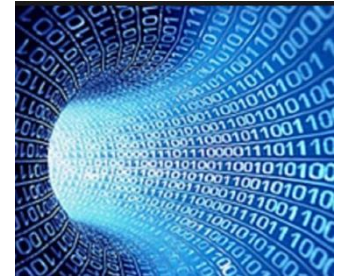
**Poor Architectural
Design &
Integration**



**No DevOps
Practices for
Scalability &
Testing**

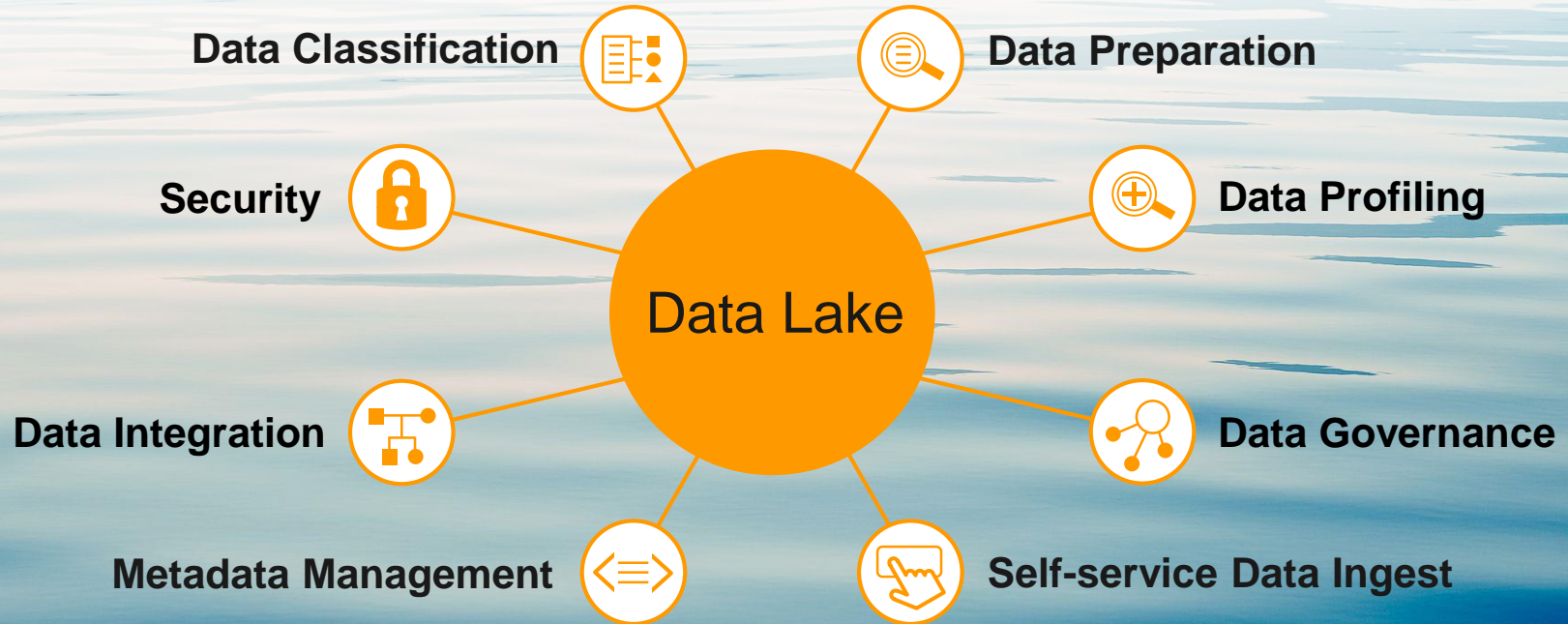


**Siloed
Operating Model**

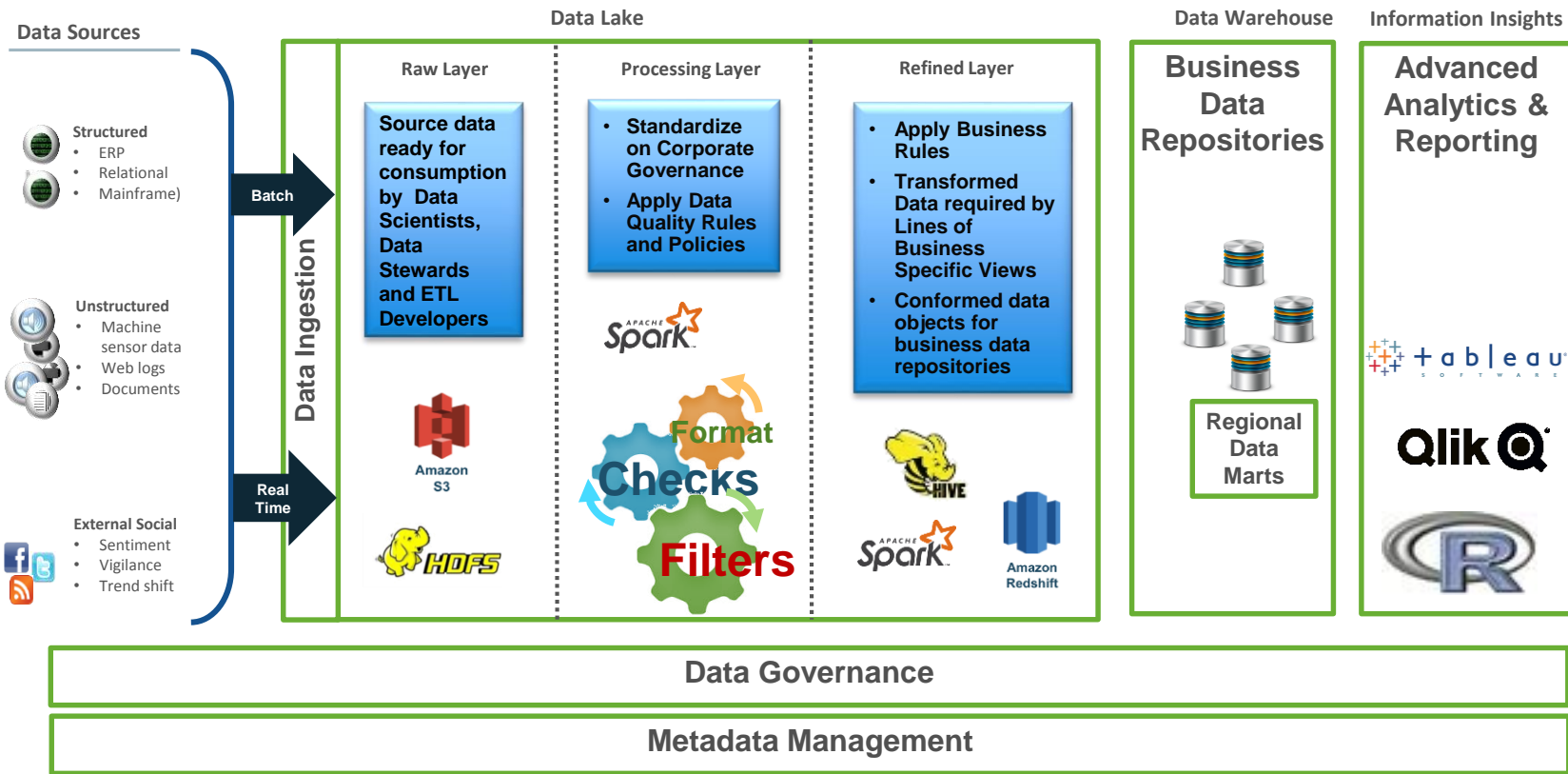


**Poor Data
Governance**

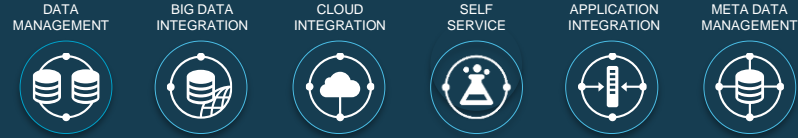
Foundational elements of operating a Data Lake



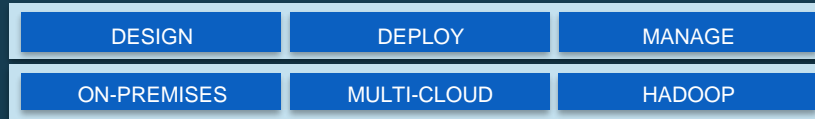
Solution Architecture of a Data Lake



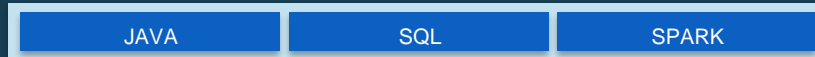
talend | Data Fabric



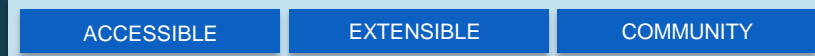
UNIFIED PLATFORM



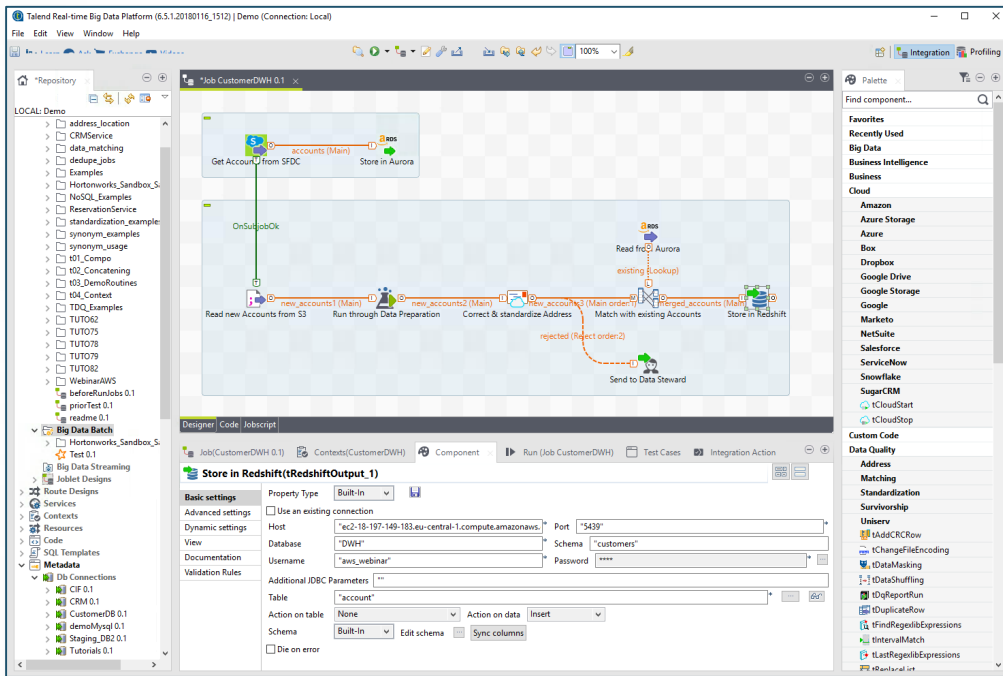
NATIVE



OPEN SOURCE



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



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 Data Prep: Self-Service for the Business Analyst

DATA PREPARATION | Help | Julien Clarysse |

WorldBeauty_SFDC_Sales Preparation | MANAGE VERSIONS | EXPORT

1 Remove fractional part on column quantity

2 Change to upper case on column Account_Name

3 Change date format on column OrderDate

4 Extract date parts on column OrderDate

5 Mask data (obfuscation) on column LastName

6 Mask data (obfuscation) on column Email

7 Lookup done with dataset S3 CASRN Chemical Open Data List. Join has been set between Product2_ExternalId and field0. The columns field3 and field4 have been added.

8 Rename column on column field4

9 Rename column on column field3

Filters 156/156

Add a filter ...

	ID	OrderItemNumber	quantity	ListPrice	Order_Effective...	Order_T
	text	integer	integer	decimal	date	
1	80258000005ospuAAA	0000000244	7	785.09	2017-10-23	
2	80258000005ospvAAA	0000000245	6	219.73	2017-10-23	
3	80258000005ospwAAA	0000000246	8	219.73	2017-10-23	
4	80258000005ospXAAA	0000000247	8	321.69	2017-10-23	
5	80258000005ospyAAA	0000000248	8	219.73	2017-10-23	
6	80258000005ospzAAA	0000000249	4	321.69	2017-10-23	
7	80258000005osq0AAA	0000000250	8	851.69	2017-10-23	
8	80258000005osq1AAA	0000000251	4	321.69	2017-10-23	
9	80258000005osq2AAA	0000000252	7	321.69	2017-10-23	
10	80258000005osq3AAA	0000000253	10	910.49	2017-10-23	
11	80258000005osq4AAA	0000000254	4	472.45	2017-10-23	
12	80258000005osq5AAA	0000000255	2	785.09	2017-10-23	
13	80258000005osq6AAA	0000000256	10	611.07	2017-10-23	
14	80258000005osq7AAA	0000000257	10	555.18	2017-10-23	
15	80258000005osq8AAA	0000000258	2	219.73	2017-10-23	
16	80258000005osq9AAA	0000000259	5	910.49	2017-10-23	
17	80258000005osqAAAQ	0000000260	9	472.45	2017-10-23	
18	80258000005osqBAAQ	0000000261	1	472.45	2017-10-23	
19	80258000005osqCAAQ	0000000262	9	555.18	2017-10-23	
20	80258000005osqDAAQ	0000000263	9	471.57	2017-10-23	

ListPrice

COLUMN ROW

Find a function ...

SUGGESTIONS

Add, multiply, subtract or divide...

Compare numbers...

Round value using halfup mode...

Remove fractional part

BOOLEAN

Negate value

CHART VALUE PATTERN ADVANCED

ROW COUNT

Occurrences

Graphical Integration Design with Talend Studio

The screenshot displays the Talend Studio interface for designing a data integration job. The main workspace shows a flowchart with the following components and connections:

- Get Account from SFDC** (Source) connects to **accounts (Main)** (Table).
- accounts (Main)** connects to **Store in Aurora** (Target).
- OnSubJobOk** (Event) connects to **new_accounts1 (Main)** (Table).
- new_accounts1 (Main)** connects to **new_accounts2 (Main)** (Table).
- new_accounts2 (Main)** connects to **new_accounts3 (Main order)** (Table).
- new_accounts3 (Main order)** connects to **Match with existing Accounts** (Join).
- Match with existing Accounts** connects to **merged_accounts (Main)** (Table).
- merged_accounts (Main)** connects to **Store in Redshift** (Target).
- rejected (Reject order:2)** (Event) connects to **Send to Data Steward** (Target).

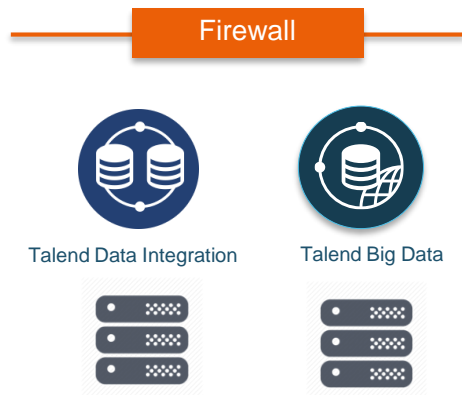
The **Store in Redshift (tRedshiftOutput_1)** component is configured with the following settings:

Property Type	Value
Host	ec2-18-197-149-183.eu-central-1.compute.amazonaws.com
Port	5439
Database	DWH
Schema	customers
Username	aws_webinar
Password	****
Table	account
Action on table	None
Action on data	Insert
Schema	Built-in
Additional JDBC Parameters	""

The interface also includes a **Repository** pane on the left, a **Palette** on the right, and a **Designer** tab at the bottom.

Flexible deployment with Talend

On-Premises



Multi-Tenant Cloud iPaaS



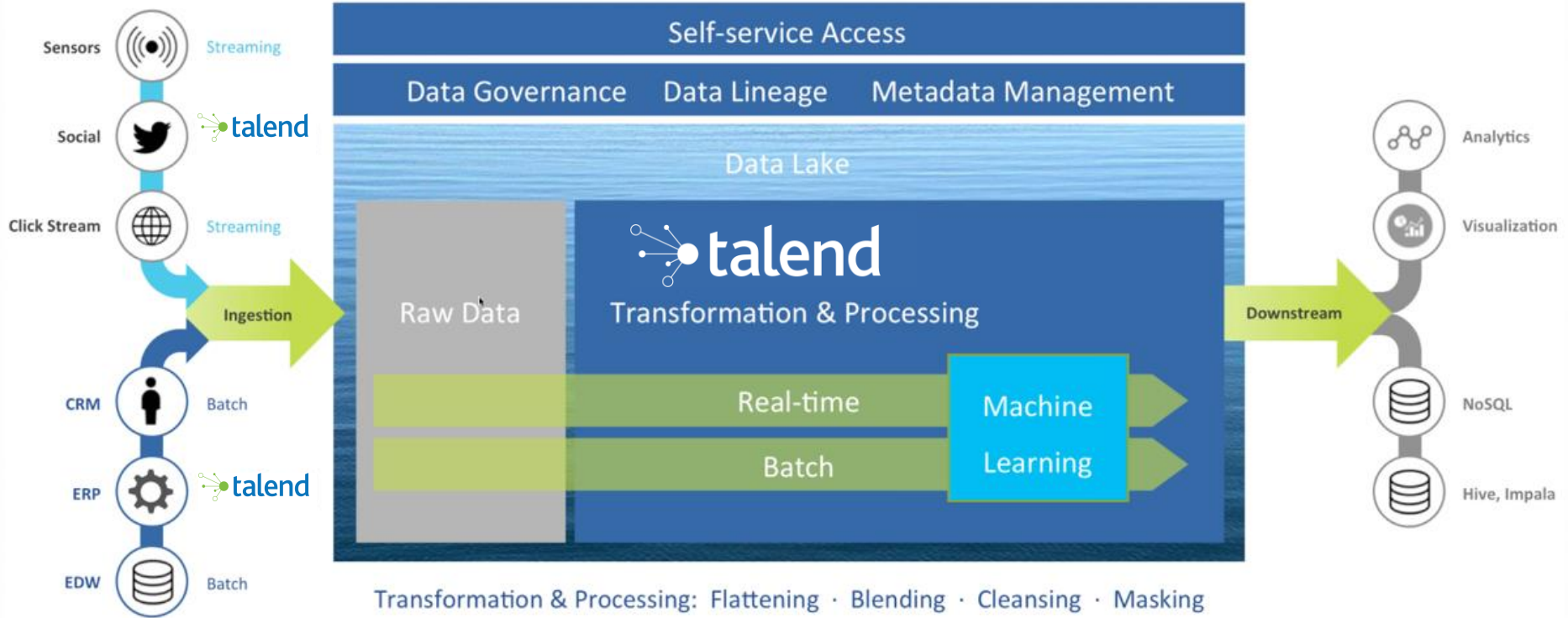
Public Cloud



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**

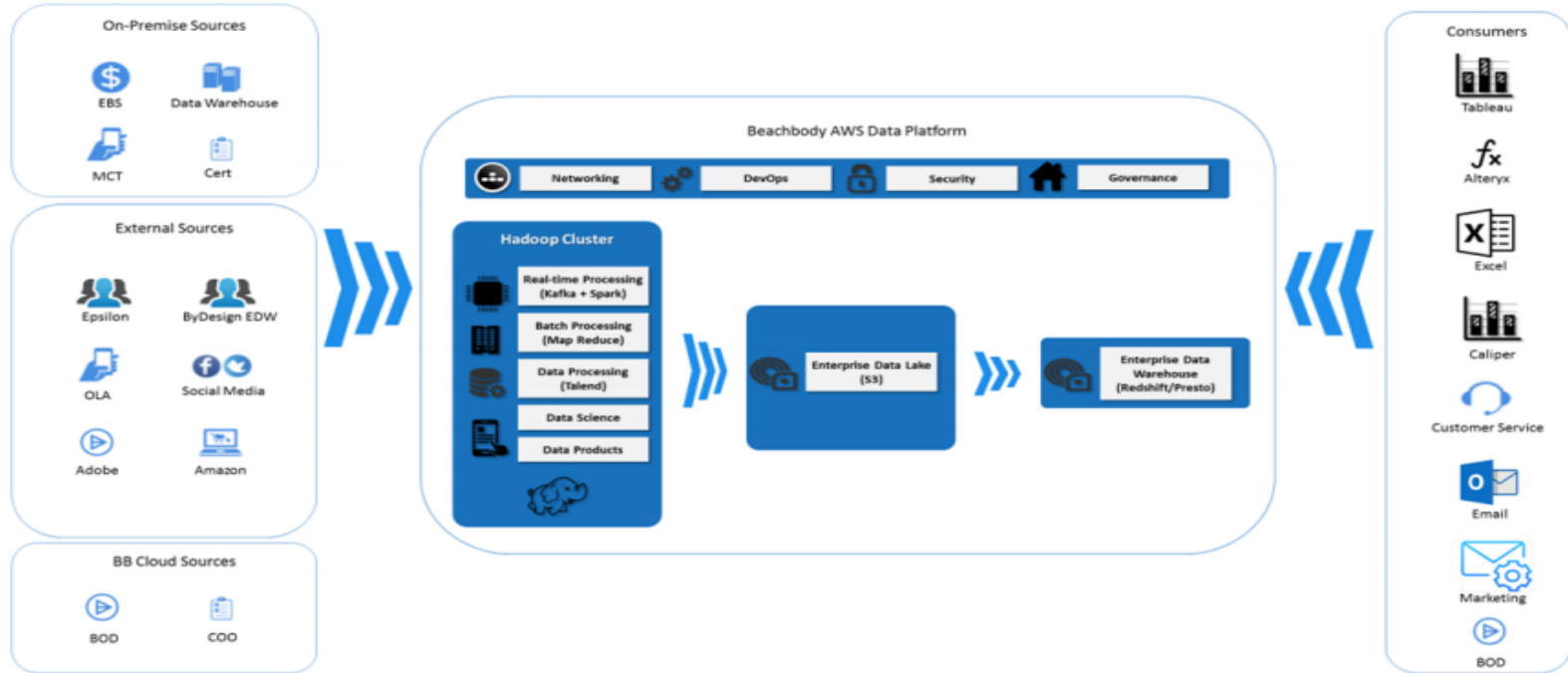


Eric Anderson, Executive Director, Data, Beachbody LLC

The Challenge - Do More Better, Faster, Cheaper

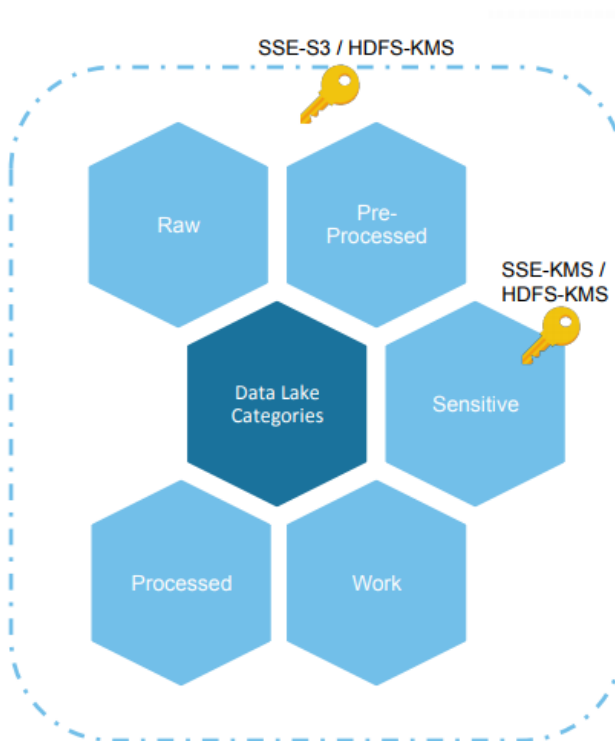


The Technology



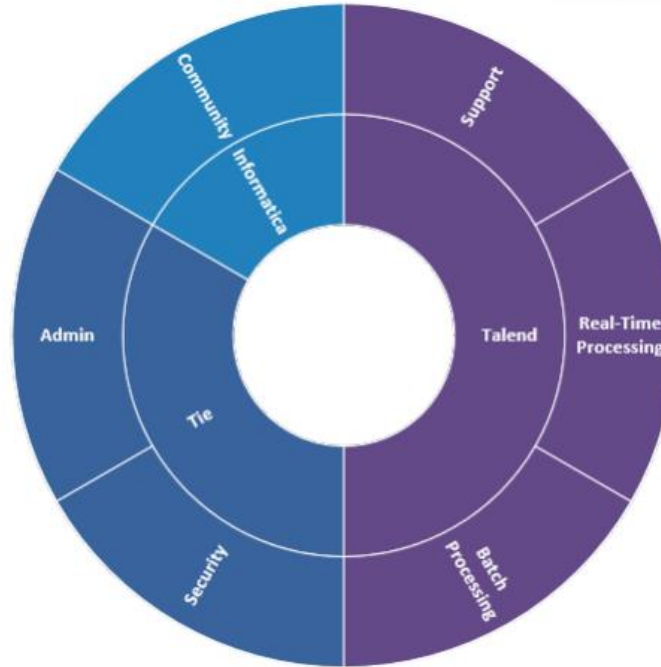
Data Lake Component - Storage

- ❑ 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

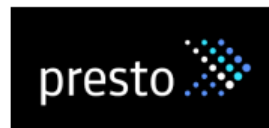
- ❑ Executes processes natively in Talend with Push-Down
- ❑ Leverages cluster security
- ❑ Will be used to:
 - ❖ Ingest & land raw data
 - ❖ Transform raw data
 - ❖ Orchestrate workflow for data science & analyses
 - ❖ Load data into RDBMS



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

BEACHBODY

- ❑ 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>