talend

# INDUSTRIALIZE DATA SCIENCE AND MACHINE LEARNING

AMINE SLIMANE – SOLUTION ENGINEERING MANAGER APAC

# AGENDA

01 TALEND PRESENTATION

02 WHAT IS MACHINE LEARNING?

03 MACHINE LEARNING ALGORITHMS

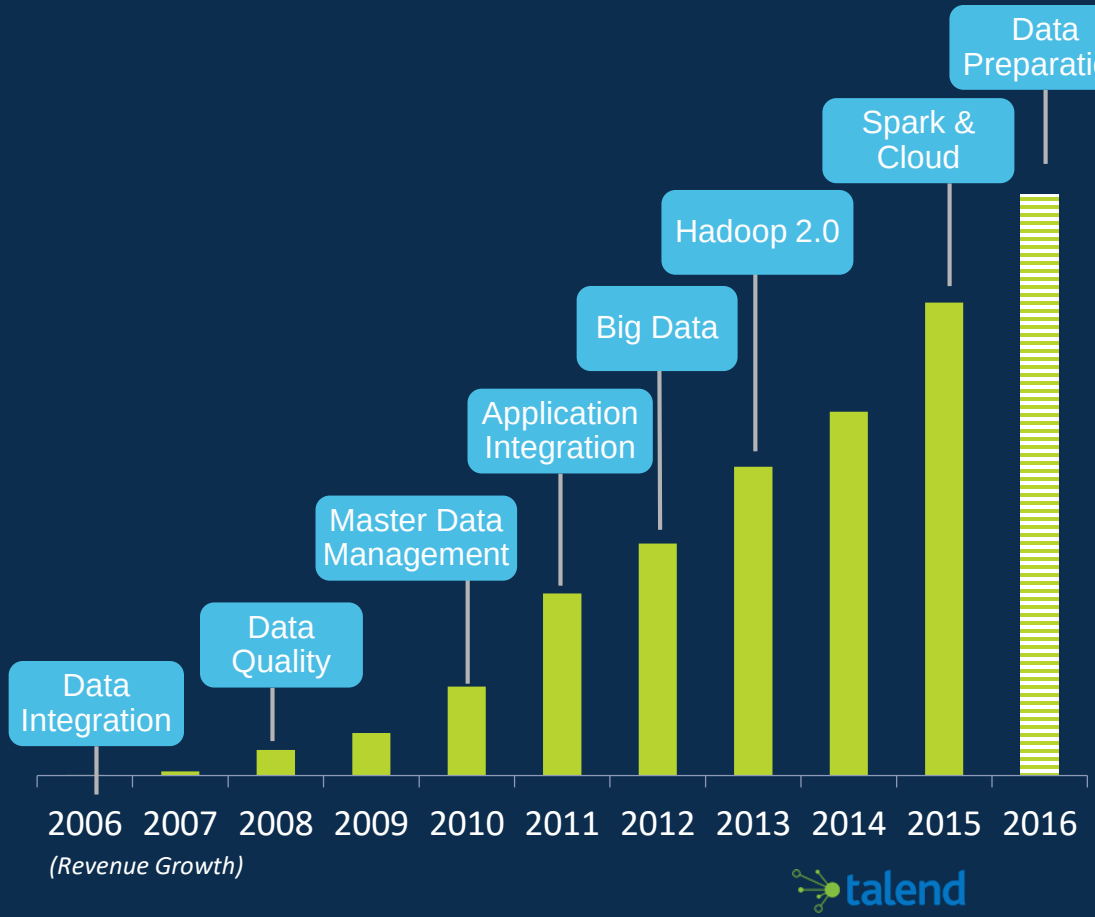
04 TALEND AND MACHINE LEARNING

05 DEMONSTRATION

# TALEND PRESENTATION

talend

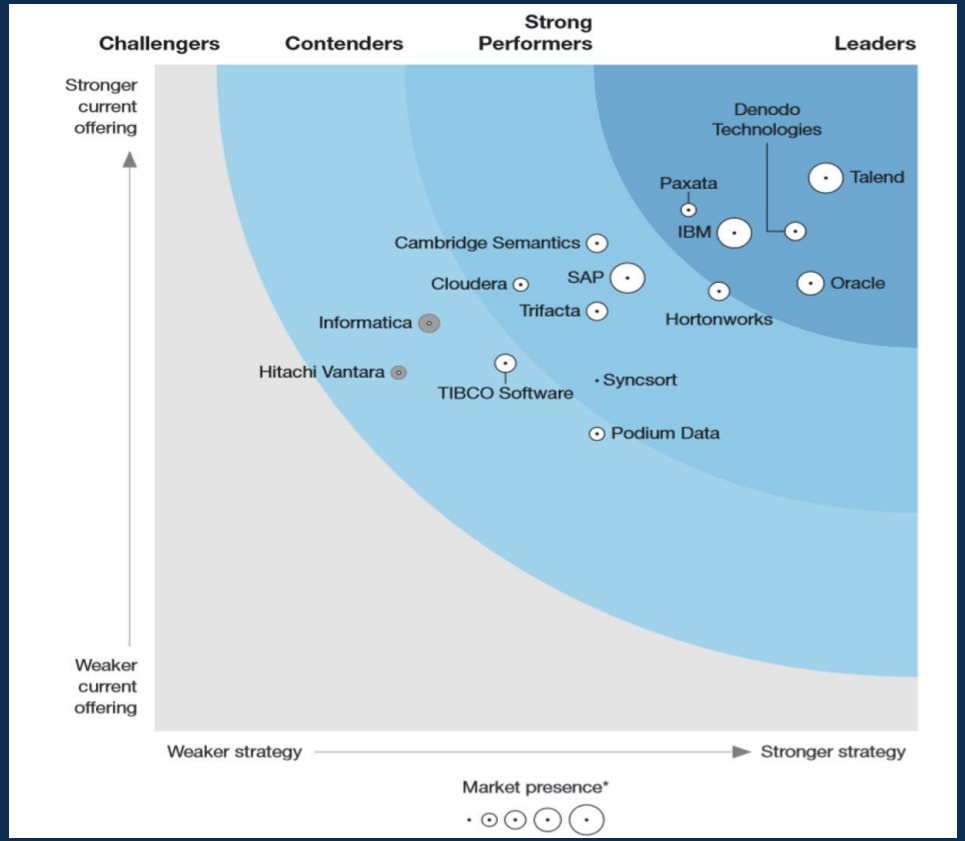
# Talend: A History of Innovation and Growth




## Key Facts

- Founded in 2006
- 1000+ employees worldwide
- 10 countries
- 1700+ customers
- 3M+ open source downloads
- 500K+ registered users
- Gartner and Forrester leader in DI, Big Data and DQ
- **Open Core Model**

# The Forrester Big Data Wave 2018

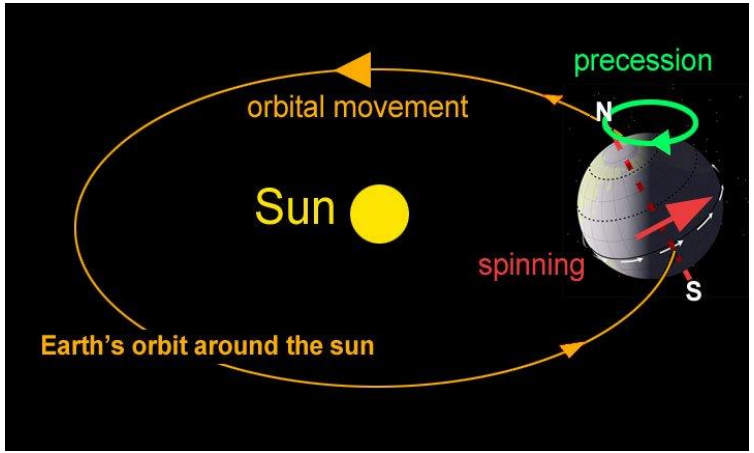


# WHAT IS MACHINE LEARNING?

talend

# WILL THE SUN RISE TOMORROW?

- How does a **machine** learn that the sun will rise every morning?
- We are **NOT** going to tell it, we want it to **LEARN**
- The machine knows **NOTHING** about how the universe works
- So, how can it learn?




talend

# SO, HOW DOES THE MACHINE LEARN?

- It simply observes and records events
- On Day 1 it sees the sun rise, but it doesn't know if it will rise tomorrow
- It uses the Bayesian interpretation of probability to work it out
- The probability the sun will rise increases each day. We just sum all the possibilities
- So, on Day 1, it will either rise or not, the probability is 0.5
- On Day 2 the probability, haven seen the sun rise once is 0.66
- At the end of the Week it is 0.857
- At the end of the Year it is 0.997

$$P[\{\text{sun rises tomorrow}\} \mid \{\text{it has risen } k \text{ times previously}\}] = \frac{\int_0^1 p^{k+1} dp}{\int_0^1 p^k dp} = \frac{k+1}{k+2}$$

# 3 MAIN TYPES OF MACHINE LEARNING

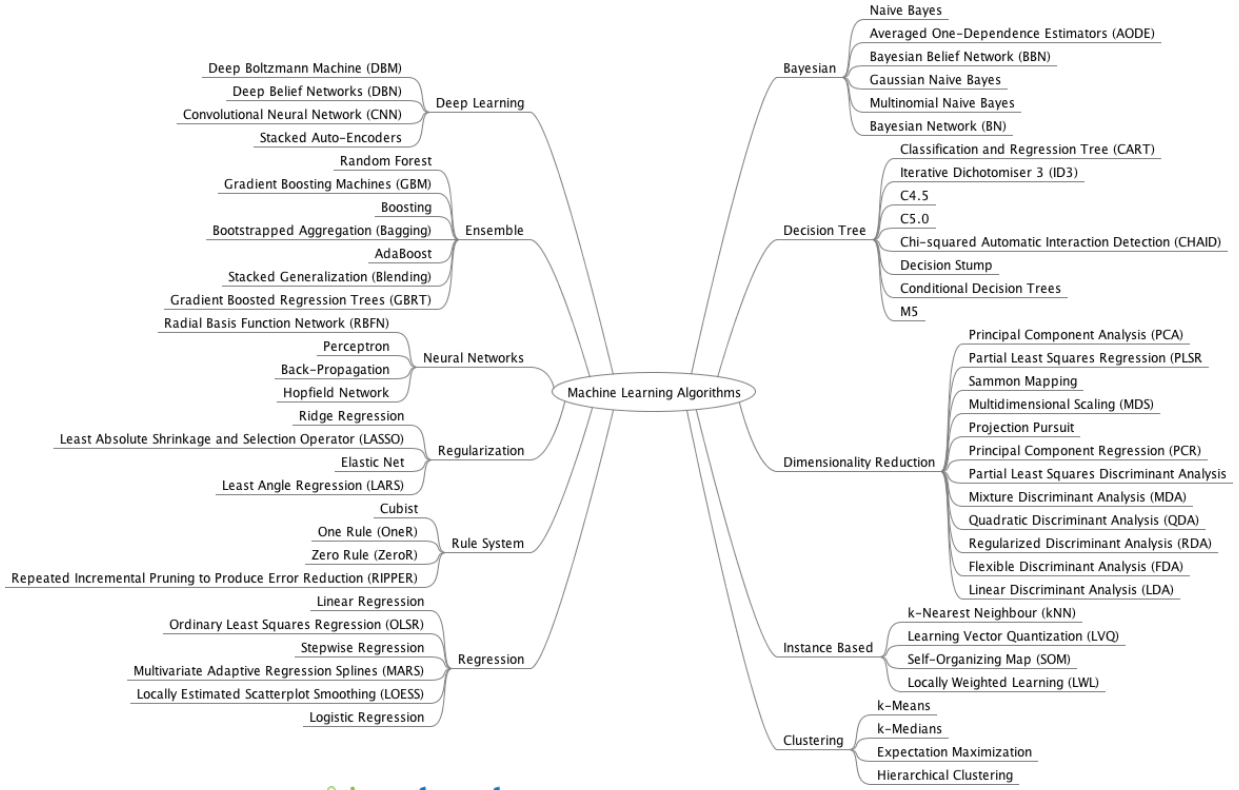
- Supervised Learning
  - A computer is presented with example inputs and desired outputs, and the goal is to learn a general rule that maps inputs to outputs
- Un-supervised Learning
  - No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning)
- Reinforcement Learning
  - A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). The program is provided feedback in terms of rewards and punishments as it navigates its problem space

# MACHINE LEARNING ALGORITHMS

# MACHINE LEARNING ALGORITHMS

*Group algorithms together into types*

*Lots of types, lots of algorithms*



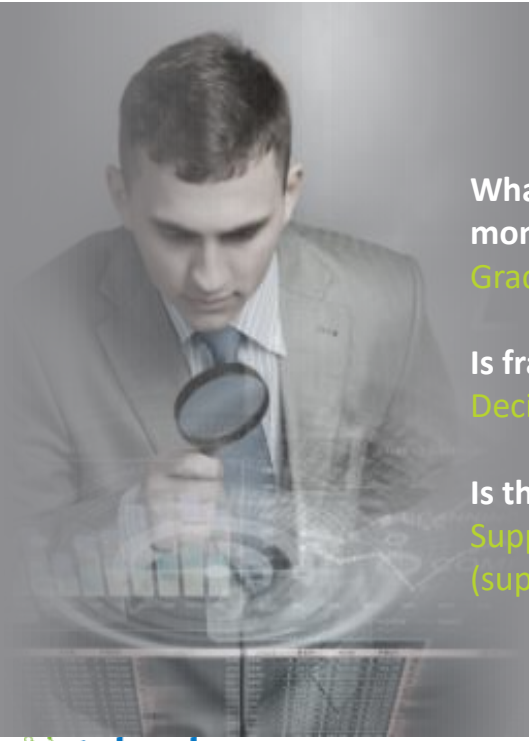
# QUESTIONS ML CAN HELP ANSWER

**Which products are likely to be bought together?**
Collaborative filtering

**Will an event happen in the future?**
Classification

**How much, what will be the number of…?**
Regression

**Who are my gold customers?**
Clustering



**What will be the price of this stock in a month?**
Gradient boosted tree

**Is fraud occurring?**
Decision tree

**Is that image a known intruder?**
Support Vector Machine
(supervised learning )

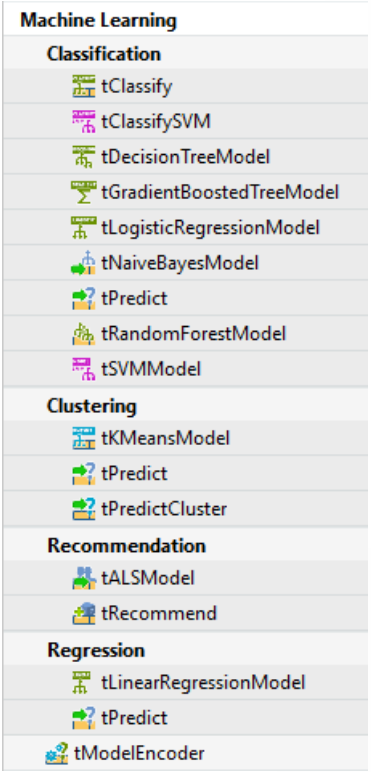
talend

# TALEND AND MACHINE LEARNING

talend

# WHY TALEND FOR MACHINE LEARNING?

- **Reduce** Machine Learning **complexities** using Talend
- Point and Click tools to **apply** various ML **techniques**
  - Classification
  - Clustering
  - Recommendation
  - Regression
- Leverage Spark for **scalability** and **performance**



# DEMONSTRATION

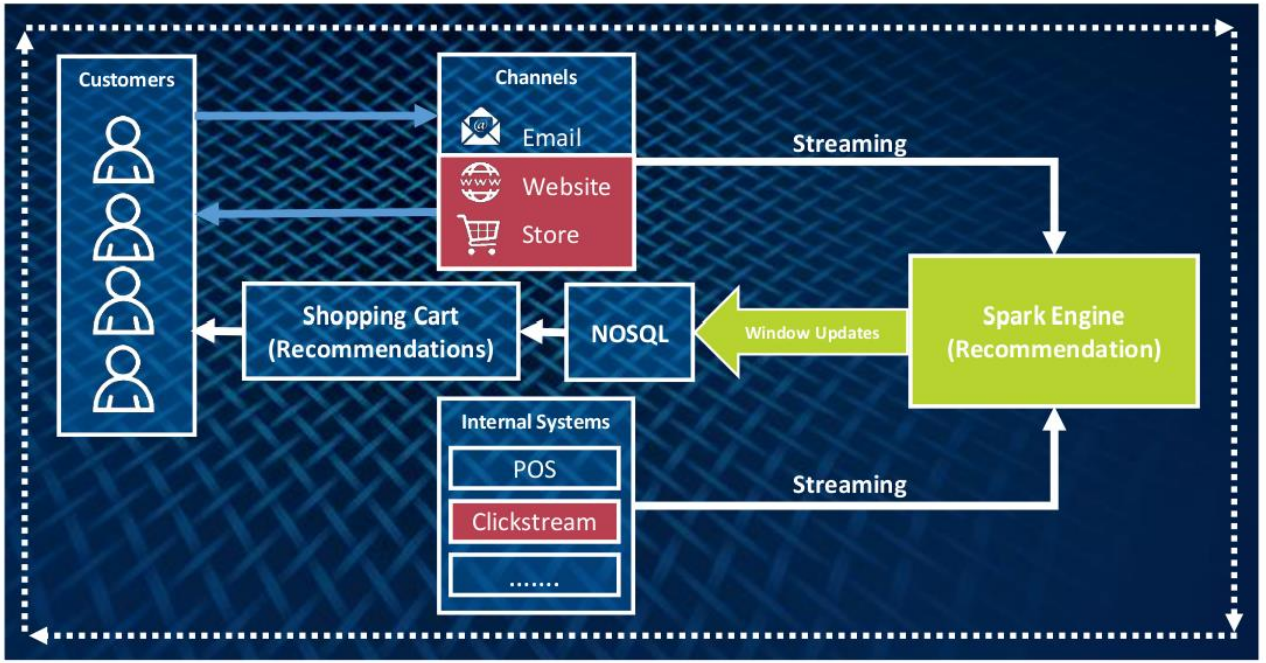
# TODAY'S GOAL

## Overview:

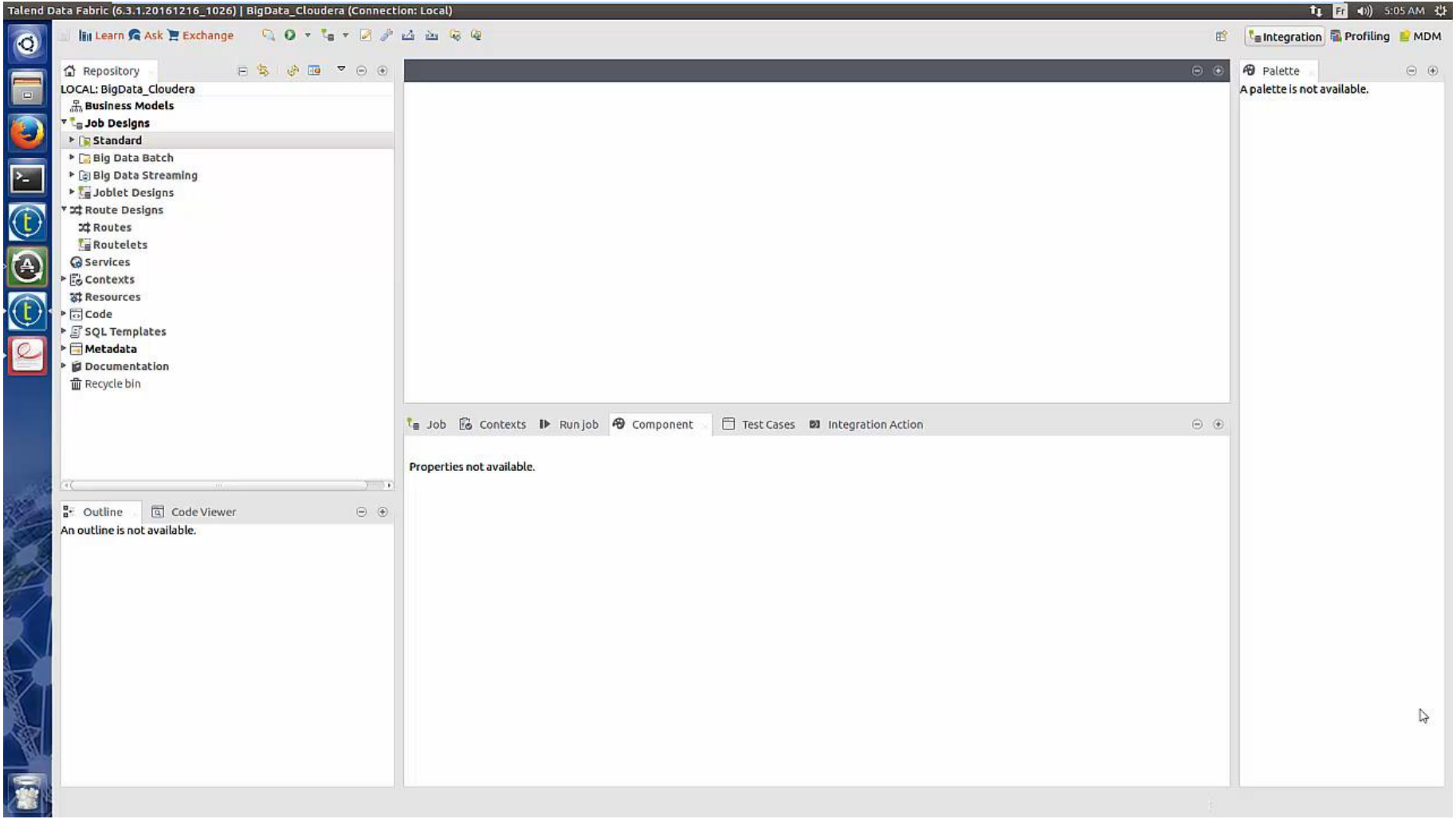
In this Demo you will see a simple version of making your website an Intelligent Application.

**You will experience:**

- Building a Spark Recommendation Model
- Setting up a new Kafka topic to help simulate live web traffic coming from Live web users browsing a retail web store.
- Most important you will see first-hand with Talend how you can take streaming data and turn it into real-time recommendations to help improve shopping cart sales.




talend


Talend Data Fabric (6.3.1.20161216_1026) | BigData_Cloudera (Connection: Local)
Learn
Ask
Exchange
Integration
Profiling
MDM
Repository
LOCAL: BigData_Cloudera
Business Models
Job Designs
Standard
Big Data Batch
Big Data Streaming
Joblet Designs
Route Designs
Routes
Routelets
Services
Contexts
Resources
Code
SQL Templates
Metadata
Documentation
Recycle bin
Palette
A palette is not available.
Job
Contexts
Run job
Component
Test Cases
Integration Action
Properties not available.
Outline
Code Viewer
An outline is not available.

# BRIDGING THE GAP BETWEEN DATA SCIENCE & IT

Data Scientist

IT

Create/tune/train predictions, models, scores

Full access to data lake for modeling

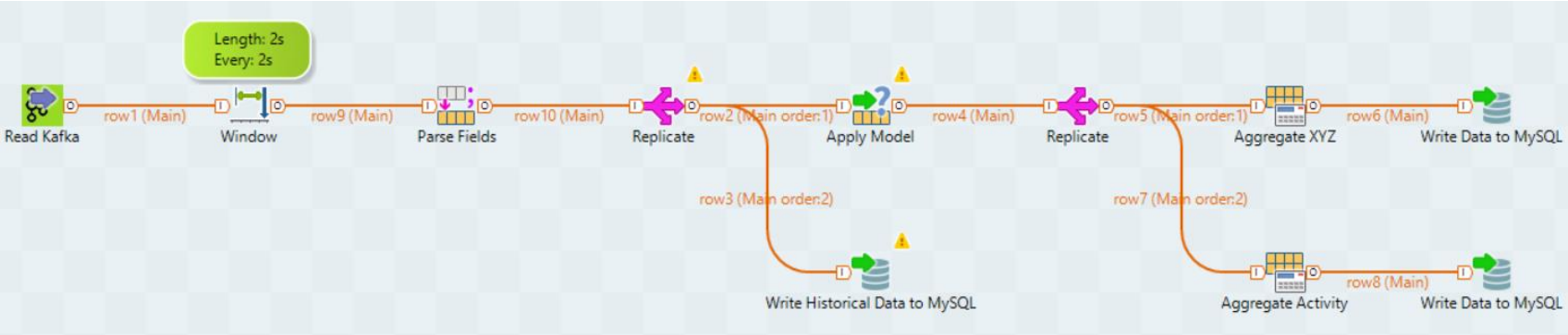
Operationalize analytics

Continuous delivery

Shorten time for IT teams to deliver

# OPERATIONALIZING THE MODEL

- Extract, Transform + Apply Machine Learning, Load
- Real-Time Processing using Spark Streaming
- Lambda Architecture (Speed + Batch Layer)
- Deploy On-Premise or In the Cloud




talend

# THANK YOU

talend


Download the Talend Sandbox at https://www.talend.com/products/sandbox