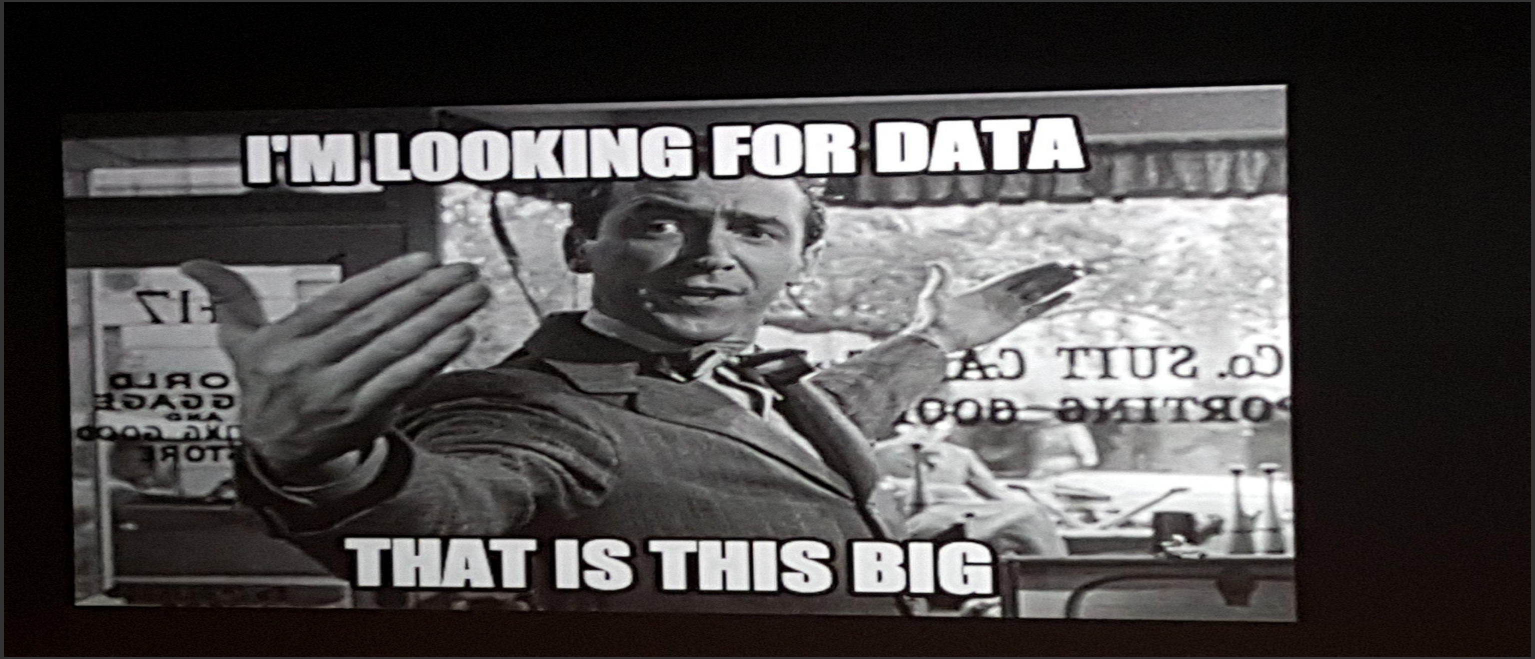
Deploying Large (Spark) ML models and scoring in near-real time @scale

Running the last mile of the Data Science journey

Subhojit Banerjee
DataScientist/DataEngineer, Founder/CTO [AbundanceAI]
@subbubanerjee
https://medium.com/@subhojit20_27731
subbu@abundanceai.com

Big Data* is not easy



Big Data* is not easy

Gartner found just 14% of companies surveyed had big data projects in production in 2015 and unchanged from the year before and slowly inching towards 15% in year 2016

**Big Data - my apologies for using the term*

**Big Data ~ Large unstructured and structured data that can't fit on a single node*

Big Data is not easy

- Cloudera: $261M in revenue, $187M in losses (down from $205M the year before, the only company to narrow its loss)
- Hortonworks: $184M in revenue, $251M in losses (up from $180M the year before)
- Alteryx: $85M in revenue, $24M in losses (up from $21M)
- Splunk: $950M in revenue, $355M in losses (up from $279M)
- Tableau: $827M in revenue, $144M in losses (up from $84M)

Big Data is not easy

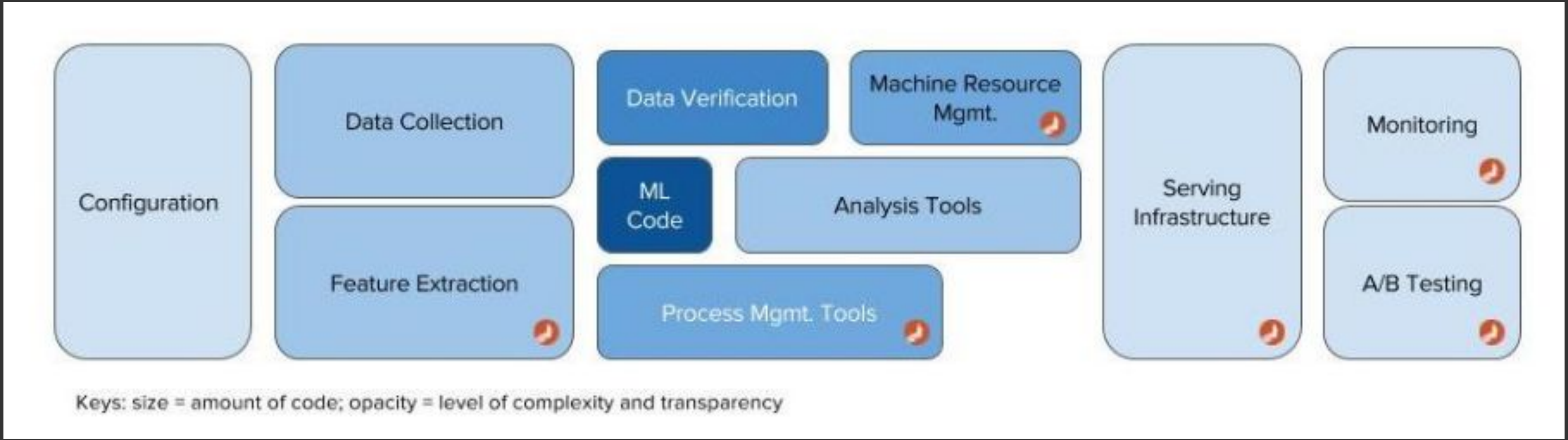
Big Data is not easy

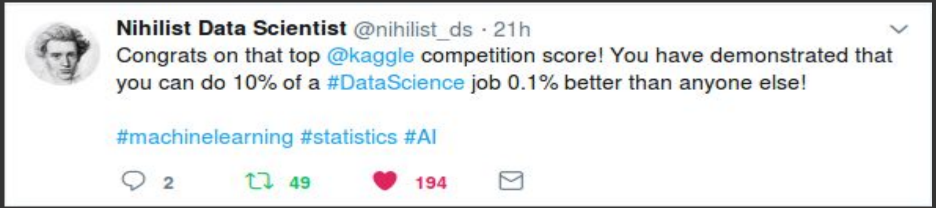
Gartner’s top obstacles for big data success were:

- Determining how to get value from Big data
- Most companies are not set up culturally or organizationally to succeed, wishing themselves agile and hoping data silos will disappear “magically”
- Technical expertise is lacking in areas like agile model development and deployment to provide quick turnarounds.

If only people were as malleable as data

Big Data is not easy




Nihilist Data Scientist @nihilist_ds · 21h

Congrats on that top @kaggle competition score! You have demonstrated that you can do 10% of a #DataScience job 0.1% better than anyone else!

#machinelearning #statistics #AI

2 49 194

** Hidden technical debt in machine learning systems - NIPS 2015*

Key Takeaways

- A better than random model has revenue generating potential from day one.Hence try to build a robust data science pipeline where models can be quickly iterated on.
- Pyspark models ***CAN** be deployed in a Scala Pipeline.
- Spark Models **CAN** be scored in “near” real time using external tools without paying the spark “distributed tax” i.e. latencies associated with spark execution plan.
- Spark Models **CAN** be dockerized and hence can leverage on best practices refined out of years of software engineering i.e. CI/CD, A/B tests etc
- And all the above can be done in a matter of minutes i.e. model creation to exposing the model @scale as an API.
- GDPR compliant features **CAN** generate models with 0.88 ROC

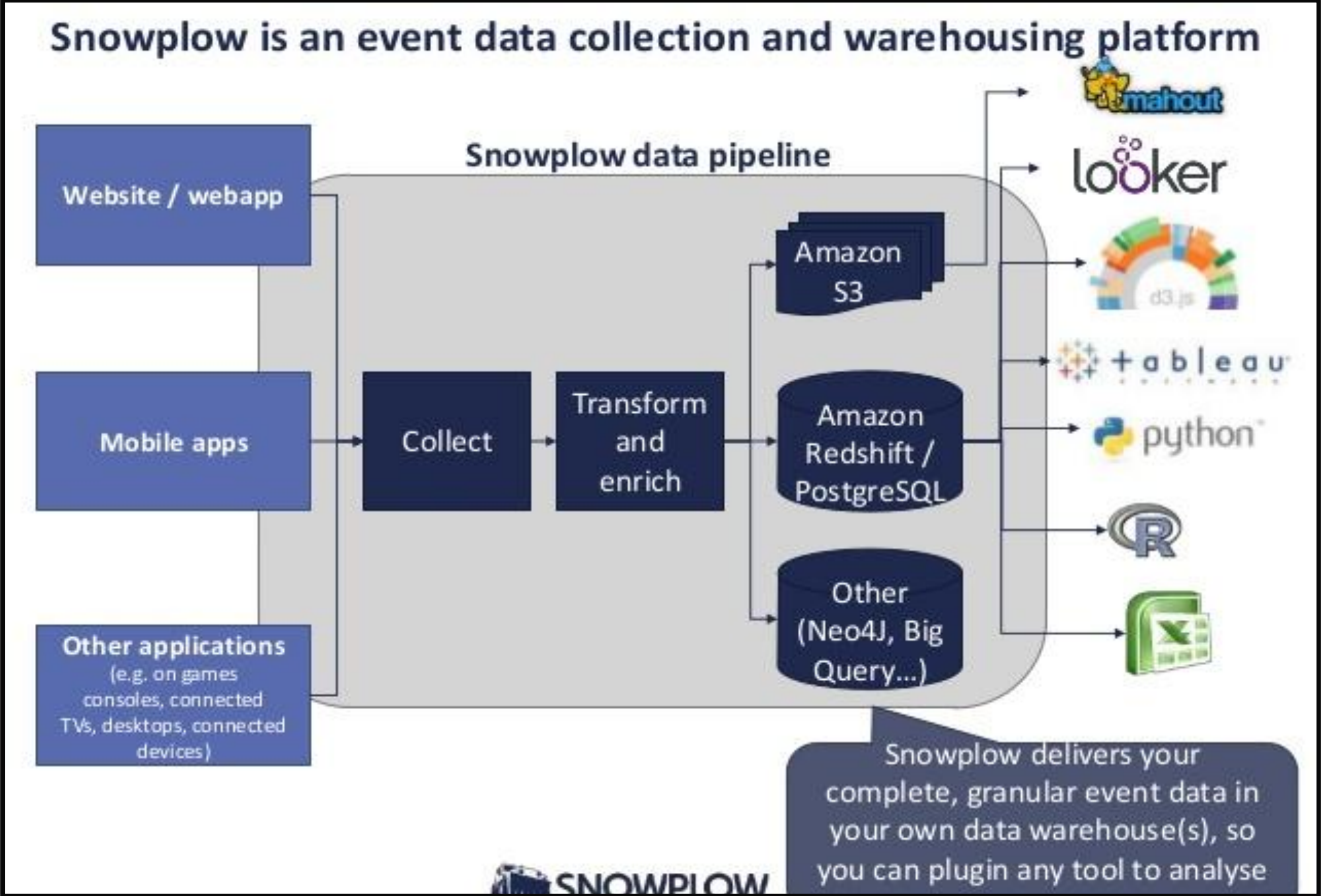
** vanilla spark has model persistence available since 2.0.0*

Business use case

Is real time segmentation of the user into buy vs defer clusters using GDPR compliant features possible on the website?

Business use case

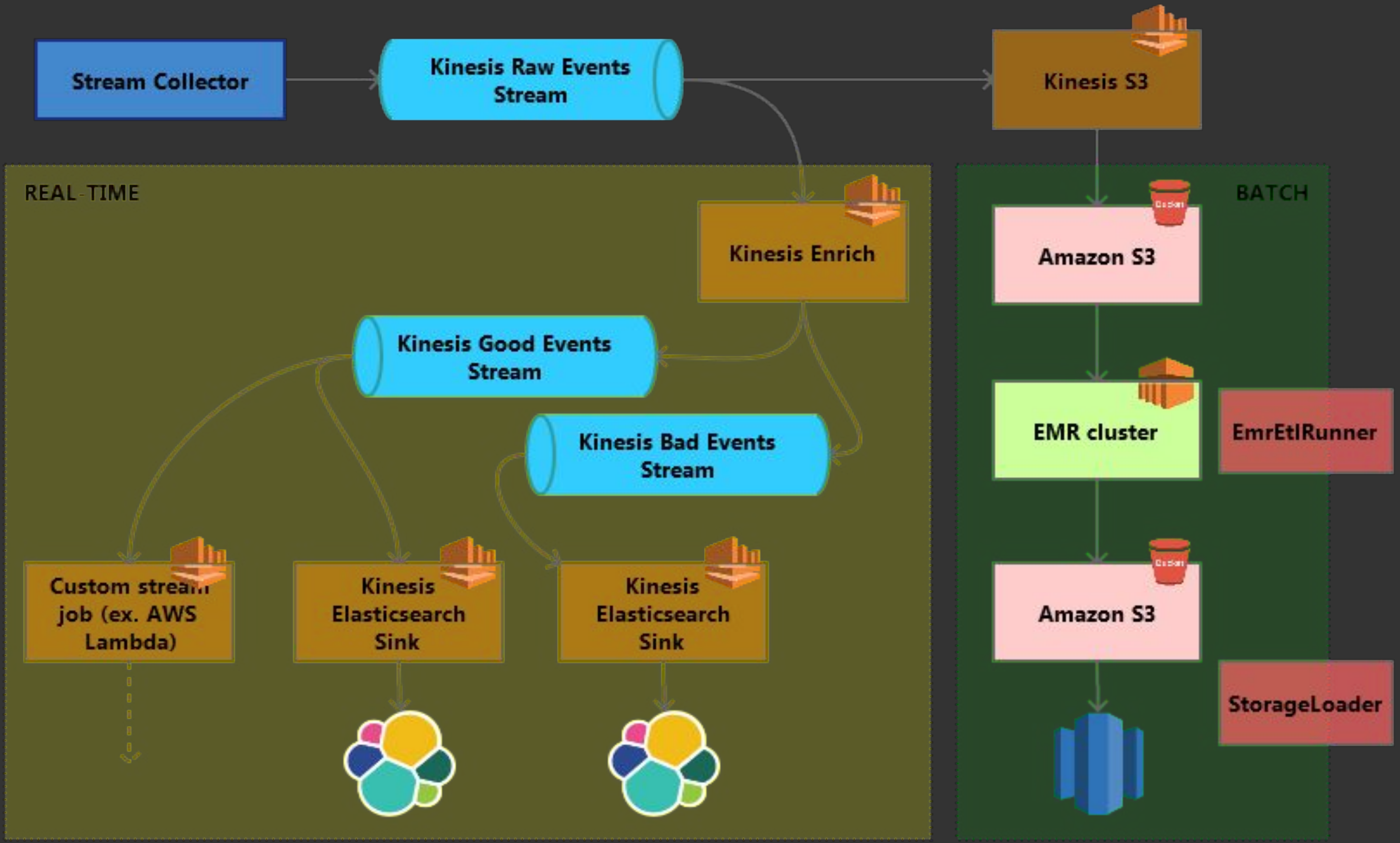
But first, we need to collect data to act on it in real time



Snowplow is an event data collection and warehousing platform

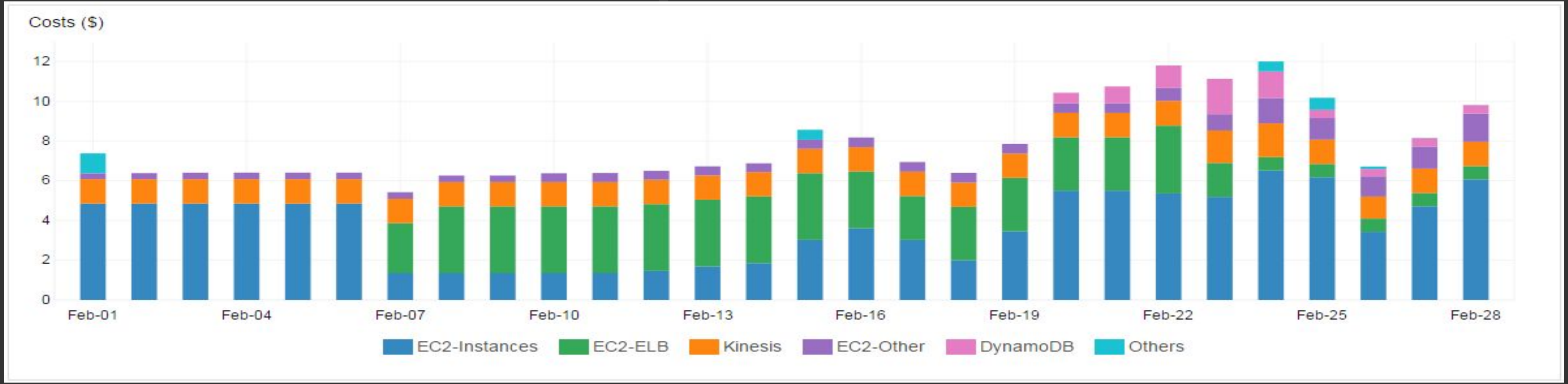
Snowplow data pipeline
Website / webapp
Mobile apps
Other applications (e.g. on games consoles, connected TVs, desktops, connected devices)
Collect
Transform and enrich
Amazon S3
Amazon Redshift / PostgreSQL
Other (Neo4J, Big Query…)
mahout
looker
d3.js
tableau
python
R
SNOWPLOW

Snowplow delivers your complete, granular event data in your own data warehouse(s), so you can plugin any tool to analyse

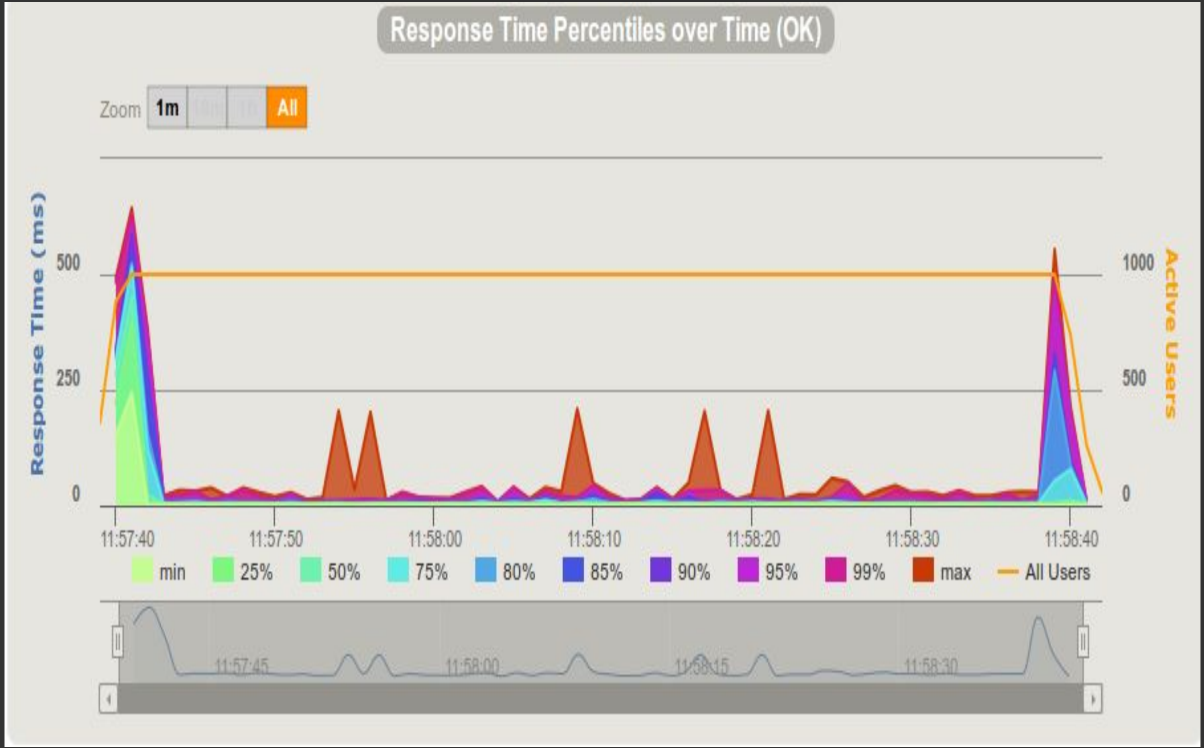

Stream Collector
Kinesis Raw Events Stream
Kinesis S3
REAL-TIME
Kinesis Enrich
Kinesis Good Events Stream
Kinesis Bad Events Stream
Custom stream job (ex. AWS Lambda)
Kinesis Elasticsearch Sink
Kinesis Elasticsearch Sink
BATCH
Amazon S3
EMR cluster
EmrEtlRunner
Amazon S3
StorageLoader

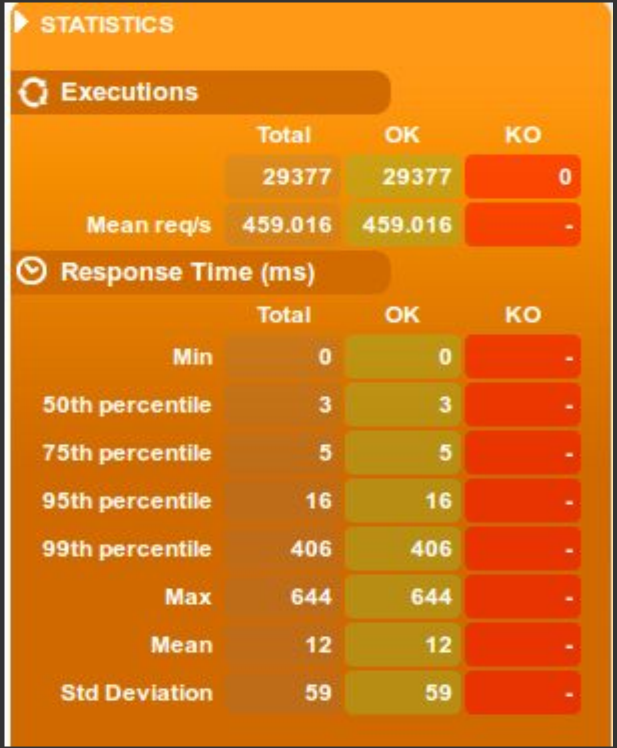
Production numbers

- ~ 23 million events per day
- ~ 140 G of granular event data collected
- cost of 6 - 12 euro per day



First solution





STATISTICS

Executions

	Total	OK	KO
	29377	29377	0
Mean req/s	459.016	459.016	-

Response Time (ms)

	Total	OK	KO
Min	0	0	-
50th percentile	3	3	-
75th percentile	5	5	-
95th percentile	16	16	-
99th percentile	406	406	-
Max	644	644	-
Mean	12	12	-
Std Deviation	59	59	-

Business use case

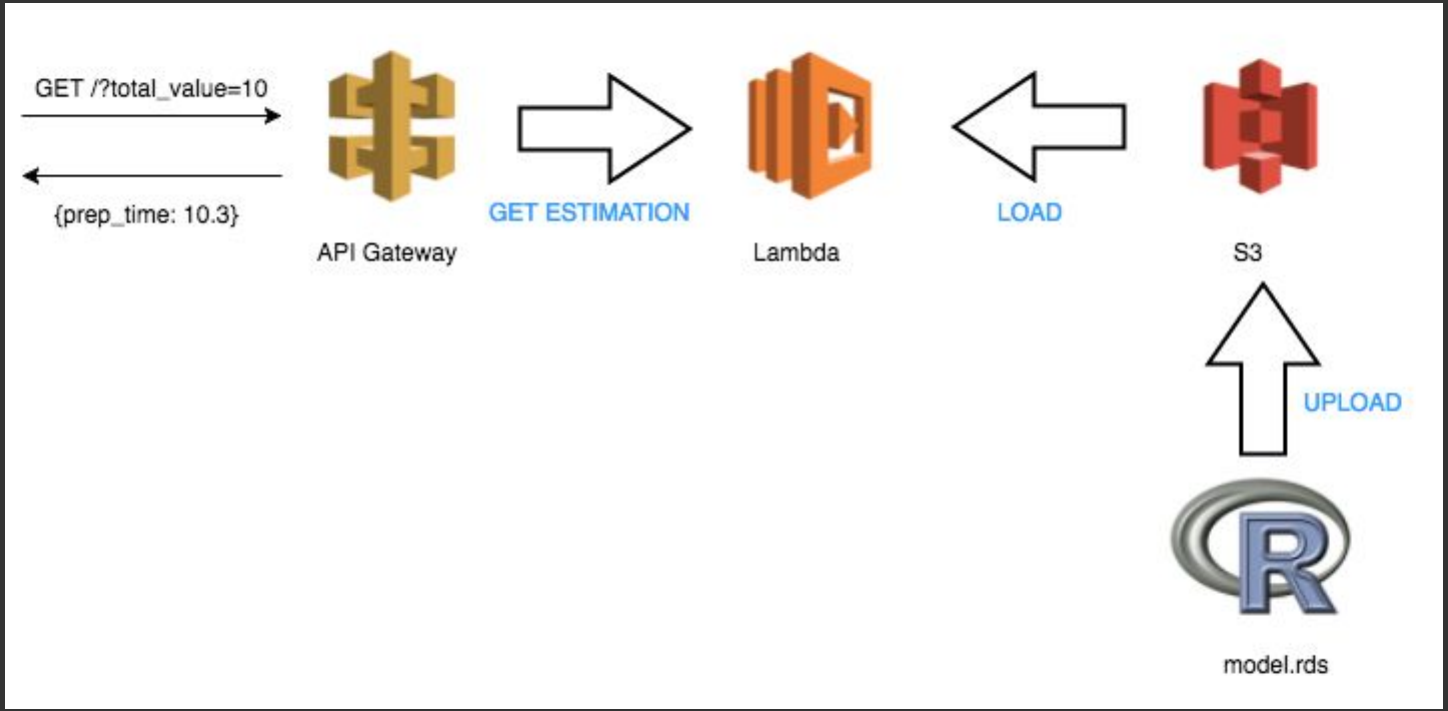
So now that we have the real time pipeline can we train and score our ML model ?

First Model

The first model was a markov chain on the sequence of webpages visited

First Solution

Aws lambda based serverless architecture



Problems faced

- Had to hack core R libraries to bring down the 212 MB ML model library to fit the 50MB compressed AWS Lambda restriction
- R is not support by AWS lambda, hence had to hack through the restriction
- Every time front end would change, our models needs to change and old data cannot be used to retrain the model.

Learnings

The effort was totally worth it as first hand one could see the scale and economies of “serverless”

“If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility... The computer utility could become the basis of a new and important industry.” -John McCarthy(1961 @ MIT Centinnial)

Better solution

Requirements:

- Decrease the time for the models to move from notebook to production
- Super Scalable - Handling Terabytes should be a cakewalk
- Has to eliminate recoding of Spark feature pipelines and models from research to production i.e. my pyspark model should be deployable into a Scala pipeline with zero or minimal code changes
- Serving/inference has to be superfast for spark models

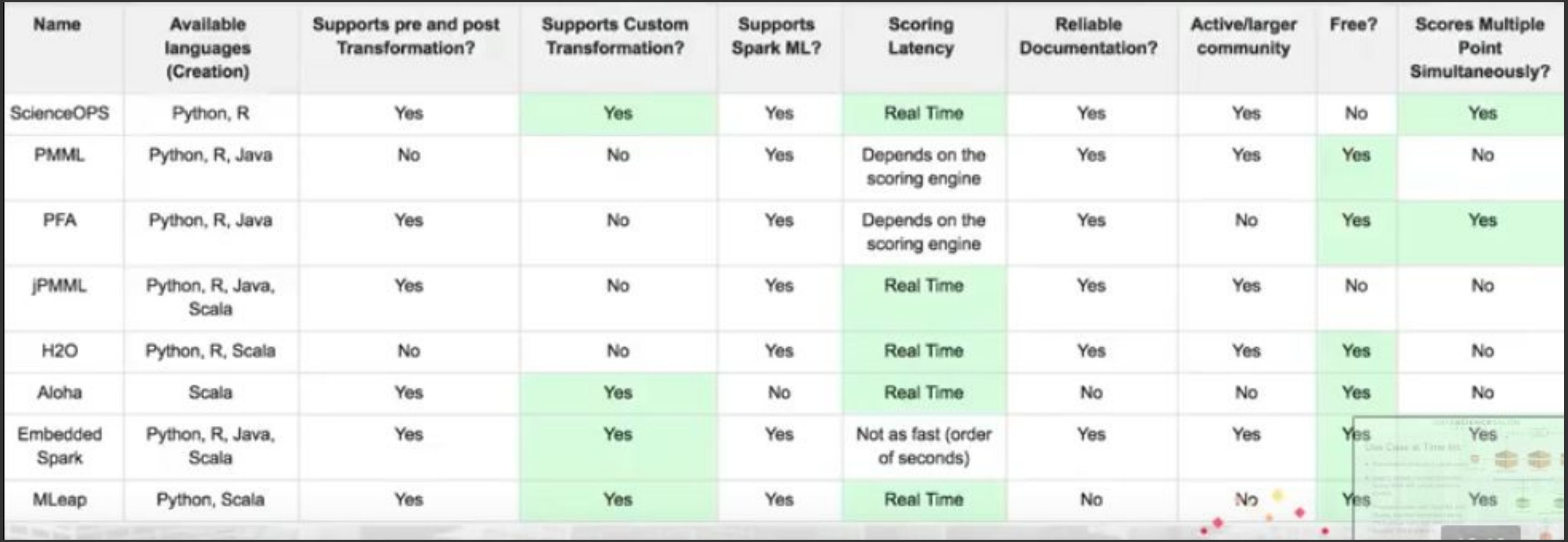
Better solution

Requirements technical analysis

- For model serving to be super fast, it was clear that inference needed to be outside the realms of Spark context.
- Has to be completely vendor neutral i.e. create a model in AWS and should be able to deploy the model in a pipeline on GCP and vice-versa.
- True test of serialization/portability: can I zip the model and send it to my coworker in an email

Why is Spark Slow in scoring

Available Options



Name	Available languages (Creation)	Supports pre and post Transformation?	Supports Custom Transformation?	Supports Spark ML?	Scoring Latency	Reliable Documentation?	Active/larger community	Free?	Scores Multiple Point Simultaneously?
ScienceOPS	Python, R	Yes	Yes	Yes	Real Time	Yes	Yes	No	Yes
PMML	Python, R, Java	No	No	Yes	Depends on the scoring engine	Yes	Yes	Yes	No
PFA	Python, R, Java	Yes	No	Yes	Depends on the scoring engine	Yes	No	Yes	Yes
jPMML	Python, R, Java, Scala	Yes	No	Yes	Real Time	Yes	Yes	No	No
H2O	Python, R, Scala	No	No	Yes	Real Time	Yes	Yes	Yes	No
Aloha	Scala	Yes	Yes	No	Real Time	No	No	Yes	No
Embedded Spark	Python, R, Java, Scala	Yes	Yes	Yes	Not as fast (order of seconds)	Yes	Yes	Yes	Yes
MLeap	Python, Scala	Yes	Yes	Yes	Real Time	No	No	Yes	Yes

Mleap

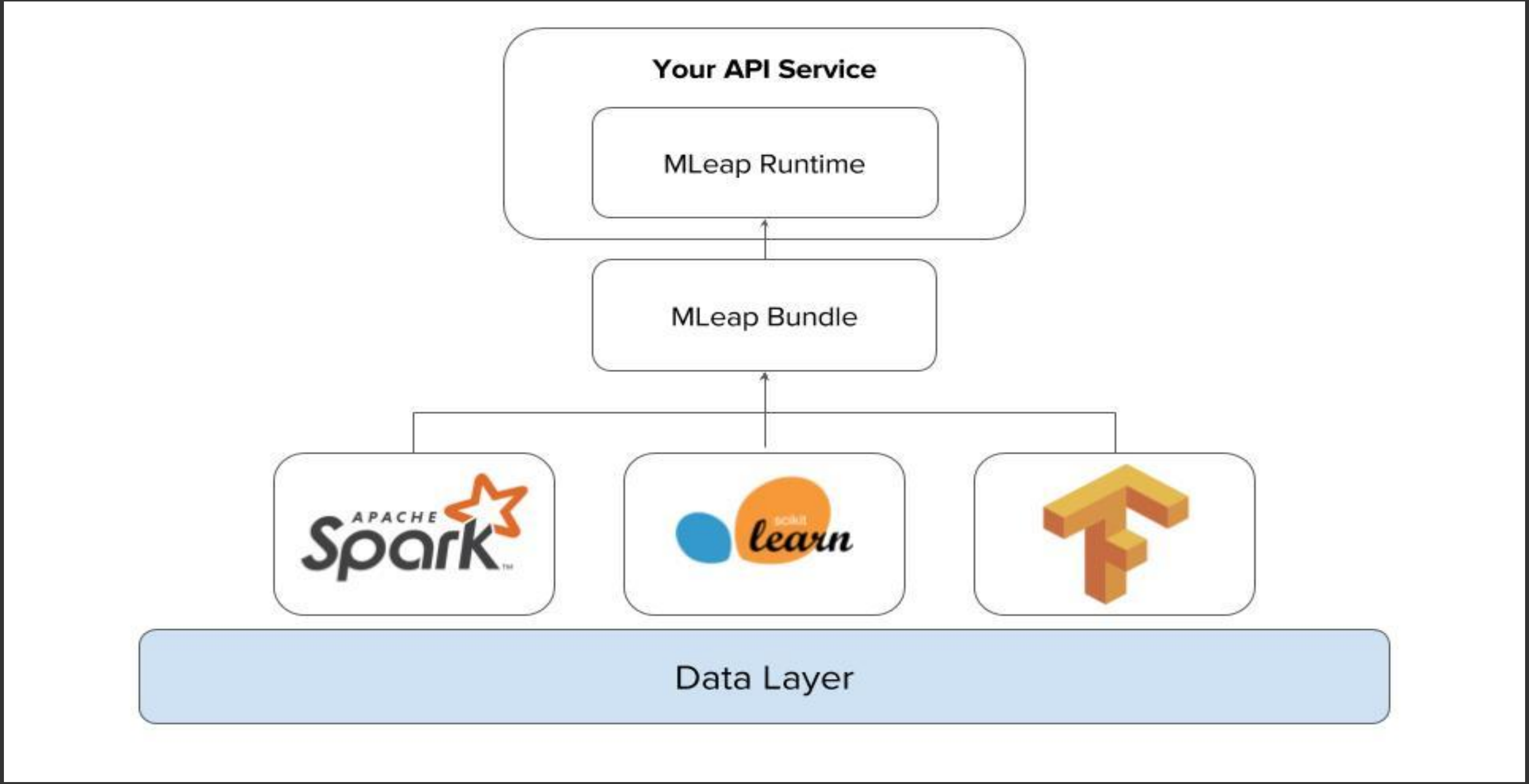
First things first: Mleap was possible due to the good work of

Hollin Wilkins and Mikhail Semeniuk

What is Mleap

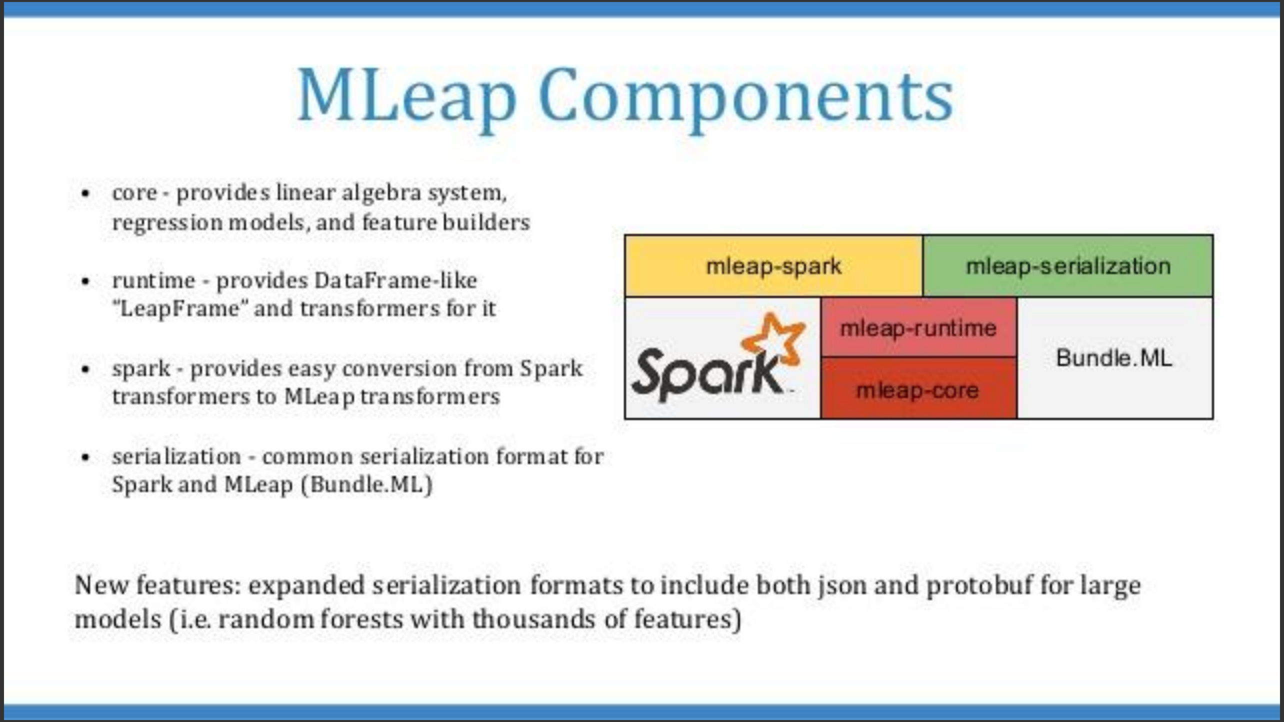
- Is a common serialization format and execution engine for machine learning pipelines.
- Supports Spark,Scikit-learn, tensorflow
- Once you serialize the models you can run it in any platform … AWS, GCP ...
- For most parts you don’t have to modify any internal code except TF
- Is completely open source, so you can open the hood.

Mleap



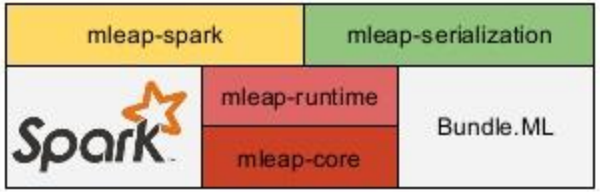
Source: Mleap docs

Mleap



MLeap Components

- core - provides linear algebra system, regression models, and feature builders
- runtime - provides DataFrame-like "LeapFrame" and transformers for it
- spark - provides easy conversion from Spark transformers to MLeap transformers
- serialization - common serialization format for Spark and MLeap (Bundle.ML)



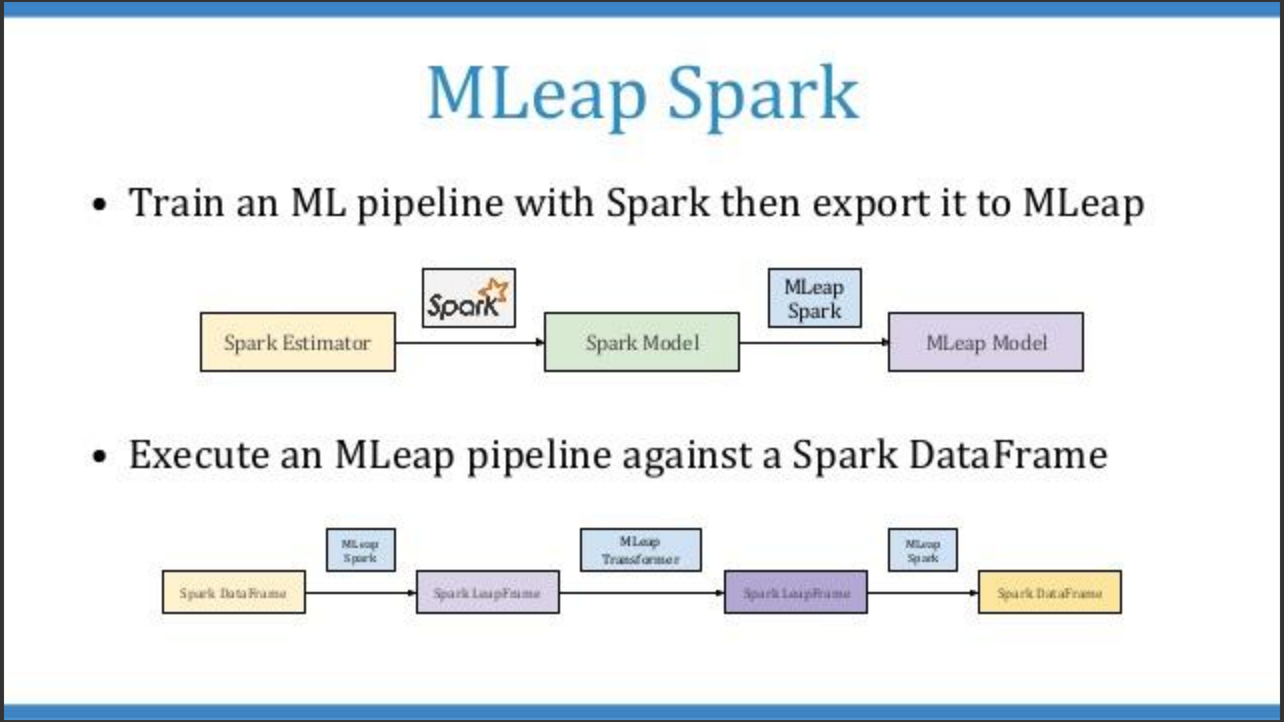
New features: expanded serialization formats to include both json and protobuf for large models (i.e. random forests with thousands of features)

Source: mleap docs

Mleap Serialization - Bundle.ml

- Provides common serialization for both Spark and Mleap
- 100% protobuf/JSON based for easy reading, compact data and portability
- Can be written to zip files and hence completely portable.

Mleap

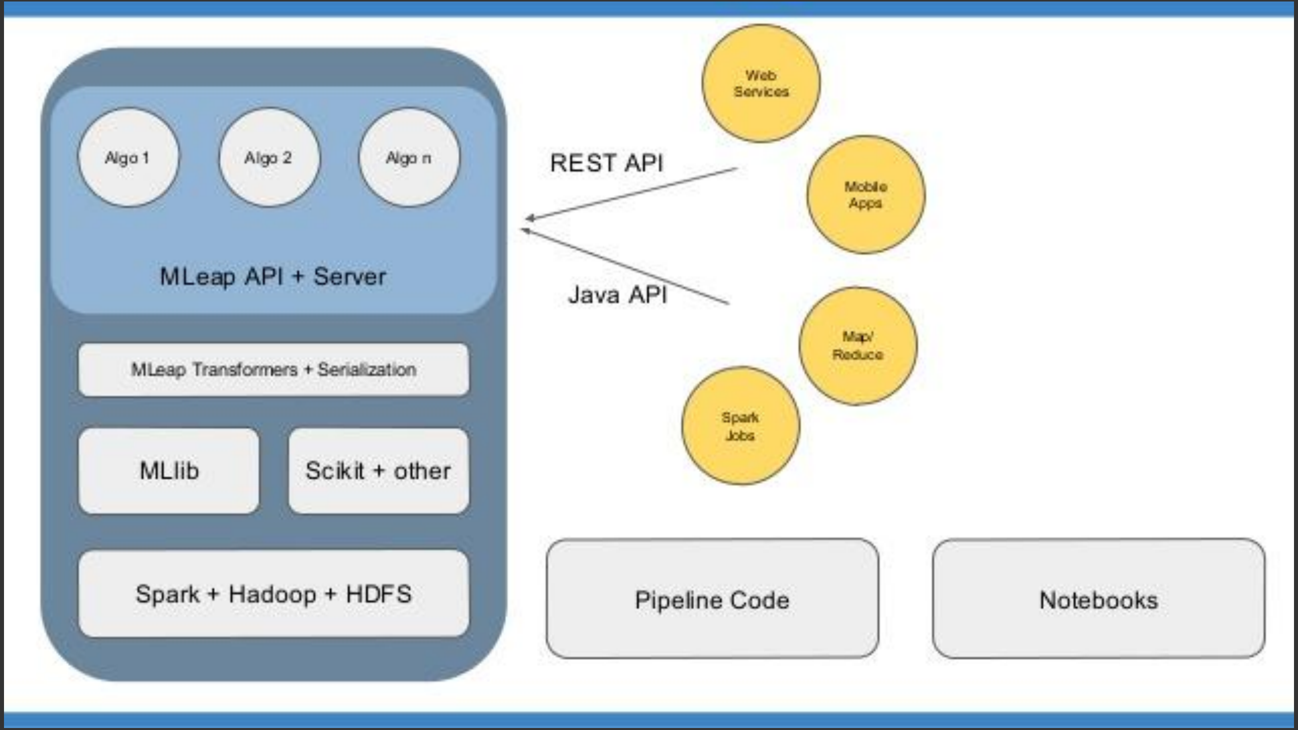


MLeap Spark

- Train an ML pipeline with Spark then export it to MLeap
- Execute an MLeap pipeline against a Spark DataFrame

Source: Mleap docs

Mleap


Algo 1
Algo 2
Algo n
MLeap API + Server
MLeap Transformers + Serialization
MLlib
Scikit + other
Spark + Hadoop + HDFS
REST API
Java API
Web Services
Mobile Apps
Map/ Reduce
Spark Jobs
Pipeline Code
Notebooks

Mleap core - Available Transformers/Estimators

Linear Algebra

- Dense/Sparse Vectors
- BLAS
- CD

Features (all)

- Vector Assembler
- String indexer
- StandardScaler
- NGram
- PCA
- MinMaxScaler
- ...

Regressors(All)

- Linear
- RandomForest
- GrandientBoosted

Clustering

- Kmeans
- GaussianMixture

Classifiers

- Logistic
- RandomForest
- GradientBoosted

Custom Transformers

- git pull

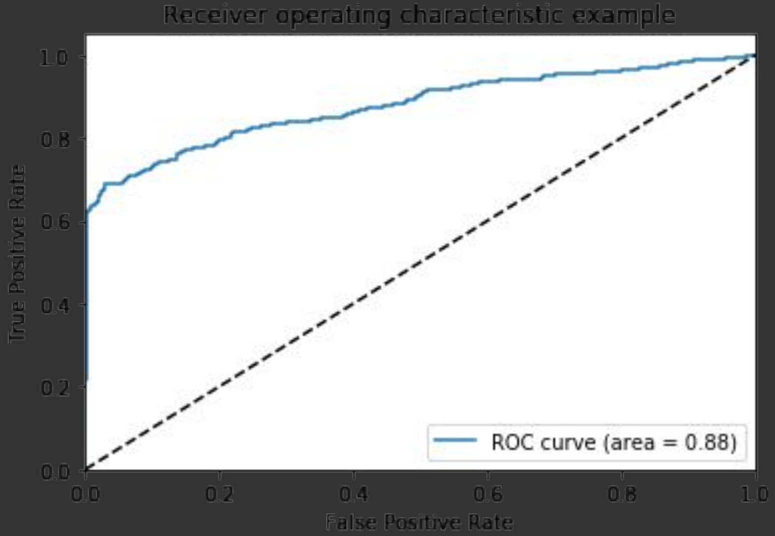
Demo

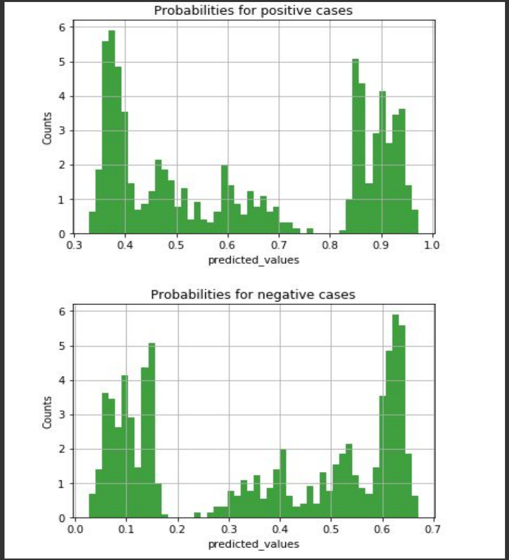
What will you see:

- Build a Pyspark model on Kaggle Data in a Jupyter notebook.
- Export the serialized model into a JSON and protobuf format (with just addition of a few libraries)
- Load the serialized pipeline into a docker container for near real model serving using a REST scala interface. (takes about 50ms for a model that spark serves in 1.5 seconds)
- Serve the Docker container from a Scalable AWS REST API in minutes

Results Achieved

Using GDPR compliant features .88 ROC and 94% on precision recall using a Pyspark model using a Single Random Forest Model (with no fancy model stacking) served in under ~50 ms trained on 100G of event data





Recsys15 Winning Method

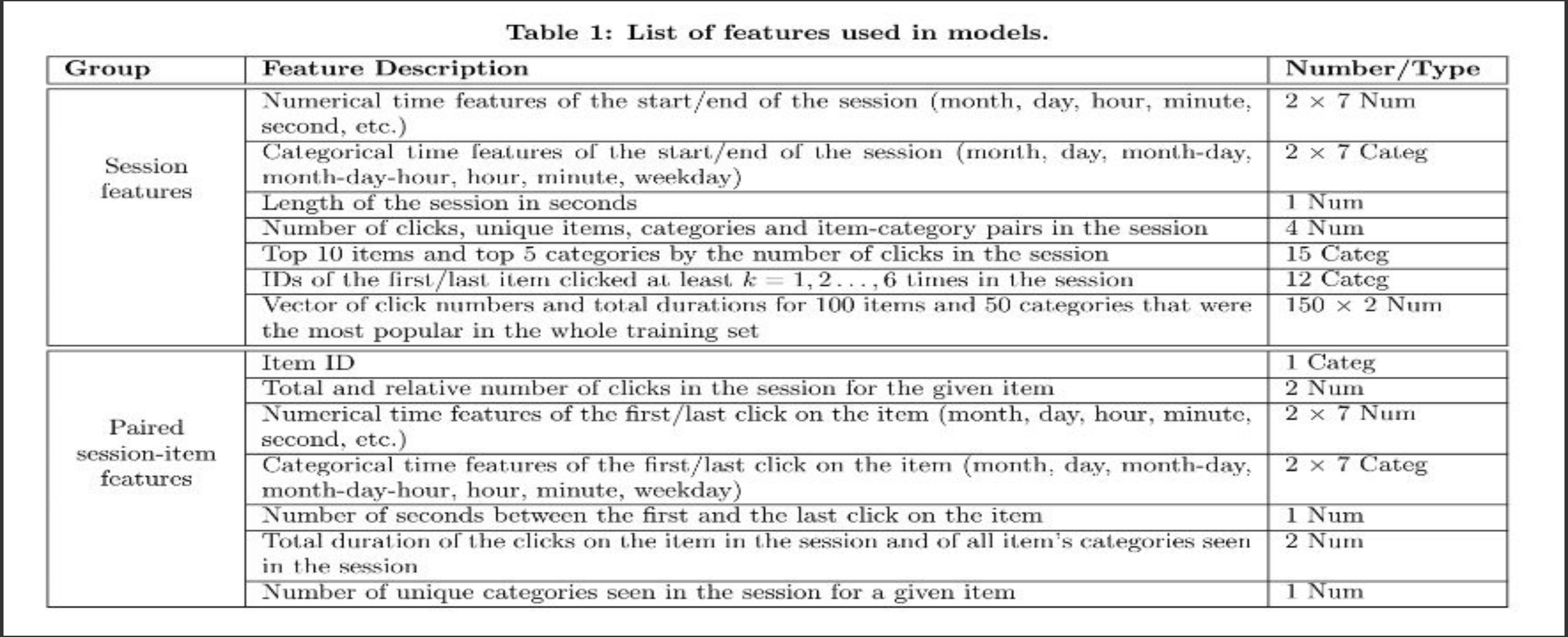
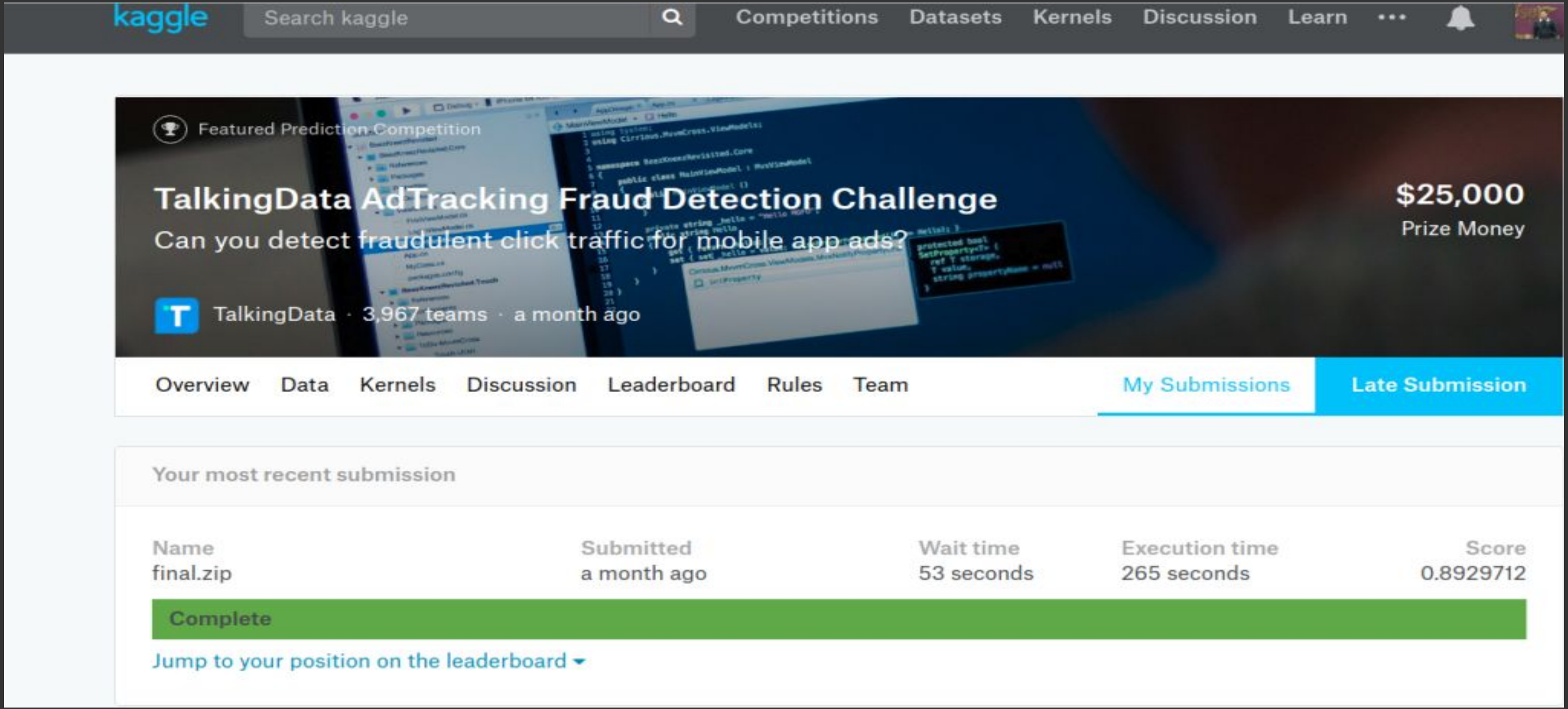


Table 1: List of features used in models.

Group	Feature Description	Number/Type
Session features	Numerical time features of the start/end of the session (month, day, hour, minute, second, etc.)	2 × 7 Num
	Categorical time features of the start/end of the session (month, day, month-day, month-day-hour, hour, minute, weekday)	2 × 7 Categ
	Length of the session in seconds	1 Num
	Number of clicks, unique items, categories and item-category pairs in the session	4 Num
	Top 10 items and top 5 categories by the number of clicks in the session	15 Categ
	IDs of the first/last item clicked at least $k = 1, 2 \ldots, 6$ times in the session	12 Categ
	Vector of click numbers and total durations for 100 items and 50 categories that were the most popular in the whole training set	150 × 2 Num
Paired session-item features	Item ID	1 Categ
	Total and relative number of clicks in the session for the given item	2 Num
	Numerical time features of the first/last click on the item (month, day, hour, minute, second, etc.)	2 × 7 Num
	Categorical time features of the first/last click on the item (month, day, month-day, month-day-hour, hour, minute, weekday)	2 × 7 Categ
	Number of seconds between the first and the last click on the item	1 Num
	Total duration of the clicks on the item in the session and of all item's categories seen in the session	2 Num
	Number of unique categories seen in the session for a given item	1 Num



Featured Prediction Competition

TalkingData AdTracking Fraud Detection Challenge

Can you detect fraudulent click traffic for mobile app ads?

TalkingData · 3,967 teams · a month ago

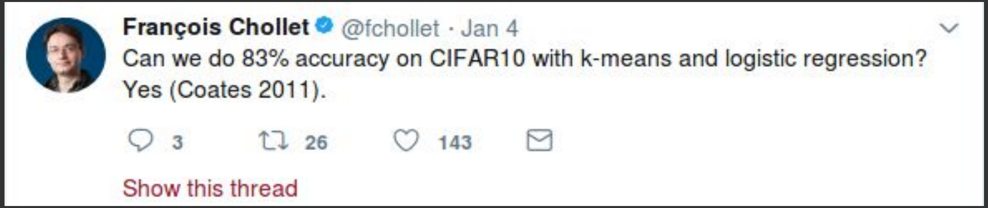
$25,000
Prize Money

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
final.zip	a month ago	53 seconds	265 seconds	0.8929712

Complete

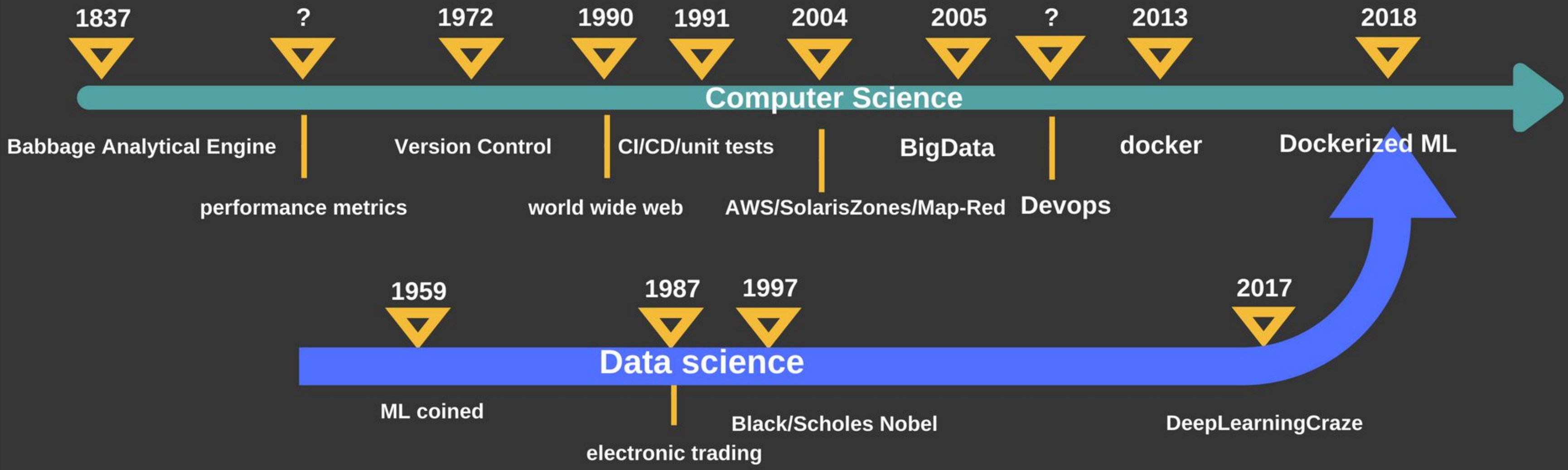
Jump to your position on the leaderboard



François Chollet @fchollet · Jan 4

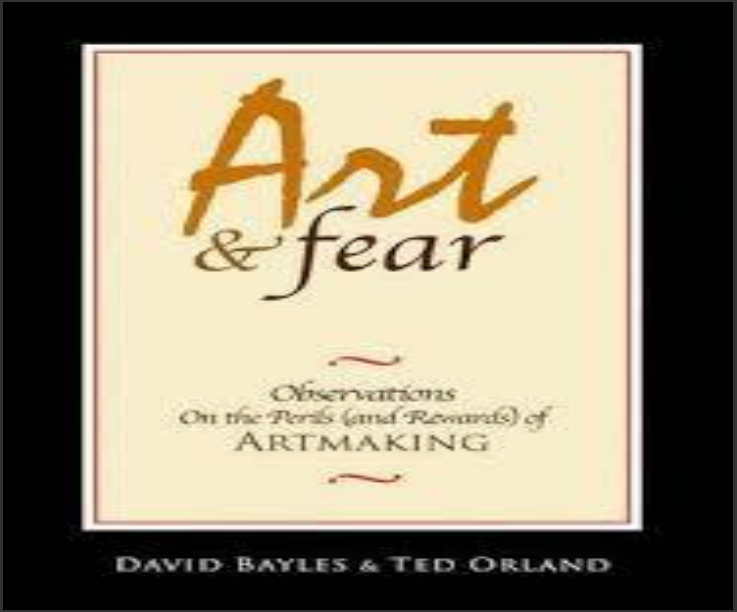
Can we do 83% accuracy on CIFAR10 with k-means and logistic regression? Yes (Coates 2011).

Conclusion



History of Controlling Electronic Chaos

Anecdote


Art & fear
Observations On the Perils (and Rewards) of ARTMAKING
DAVID BAYLES & TED ORLAND

The works of highest quality were all produced by the group being graded for quantity.