PML

Machine Learning
the PunchPlatform way

PunchPlatform team

Agenda

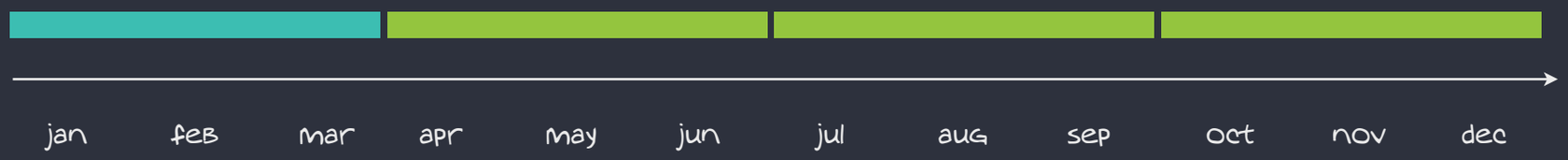
- Problem Statement
- Spark Concepts Made Easy
- From R&D to Production
- An Inside Look at the Architecture
- Thanks

Problem Statement

Problem Statement

You have a year of data into your (Punchplatform) ElasticSearch. You use it to

- react : Using Kibana and the power of ElasticSearch query language, you keep an eye to what is going on you system.
- report : compute powerful aggregated indicators overs last weeks of data. You keep control of important or suspect activities.
- investigate : perform on demand search and analysis over historical data (up to one year)
- IOCs : check for indicator of compromise over arbitrary periods of data



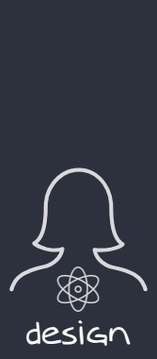
You want more.

You wan to benefit from data analytics and machine learning promises.

How do you do it ?

Spark Made Simple

PML : Spark Concepts Made Easy



Plan

Fit Every Night on last day of (Elasticsearch/ Kafka) data

Transform last hour of (ElasticSearch/Kafka) Data every half an hour

Input Data

Fit

Model

Scored Data

Transform

View

Evaluate

Alert

Detection Data

Fit Job Example

Data Input
ElasticSearch last day of firewall logs

Punch Stage
Field selection, enrichment, ..

Spark Stage
SparkQL

Spark Stage
K-Means, Regression, ...

Spark Stage
Filter

Model Output
Save the computed Model

Transform Job Example

Punch Stage
Field selection, filtering, ...

Spark Stage
(say) K-Means/Regression/...

Data Output
Save the scored data to ElasticSearch

Model Input
from Fit Job

Live Input
Real Time Firewall Logs

Spark Stage
SparkQL

Spark Stage
select fields

Leveraging Spark MLib

https://spark.apache.org/mllib/

ML Algorithms

Classification:
  logistic regression, naive Bayes,..
Regression:
  generalized linear regression, survival regression,..
Decision trees, random forests,
  gradient-boosted trees
Recommendation:
  alternating least squares (ALS)
Clustering:
  K-means, Gaussian mixtures (GMMs),..
Topic modeling:
  latent Dirichlet allocation (LDA)
Frequent itemsets, association rules,
  sequential pattern mining

ML Workflow

Feature transformations:
  standardization, normalization, hashing,..
ML Pipeline construction
Model evaluation and hyper-parameter tuning
ML persistence:
  saving and loading models and Pipelines

Utilities

Distributed linear algebra: SVD, PCA,..
Statistics: summary statistics, hypothesis testing,..

PML Benefits

Spark is the future of data analytics.
Punchplatform makes it real simple to design arbitrary Spark processings.
Using only configuration files : no coding.
Benefiting from Punchplatform data normalisation.

Design, configure, deploy, run, monitor, visualise.
Focus on your use cases, not on the technological stack.

You have no limit : you can design arbitrary Spark processing.

Be part of the Thales Data Analytics community.
Join us to design clever ML processing.

From R&D to Production

Is Applied Data Science Easy ?

Writing a POC on top of Spark on top of Azure/Amazon is easy. It takes a student and a few days. Delivering valuable and effective algorithms to production platforms is another story:

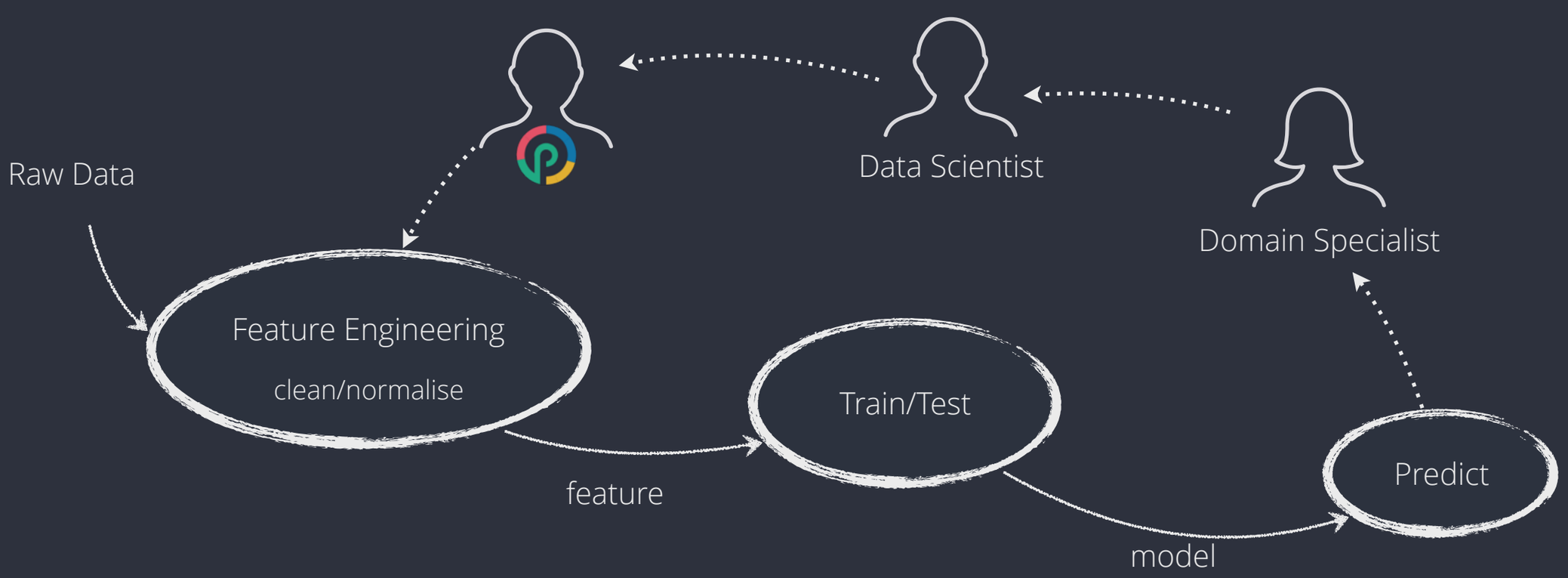
- What for ?
  - Use Case ? Who can define one ?
  - Anomaly Detection ? What is an anomaly ?
  - Learning ? Learning what ?
- Who is in charge ?
  - Data Scientist, cybersecurity or domain expert, Engineers, Geeks ?
- Assume all that is cleared up : How ?
  - Deploy/Run/Monitor
  - Where's the data
  - Is the data ready to be used ?
  - Must it be resilient ? Idempotent ? Exactly-Once

Solving the Equation

Key to success :

- small, agile, integrated team
- well defined process
- clear and shared vision of achievable steps : MVPs

..... and something like a



How It Works

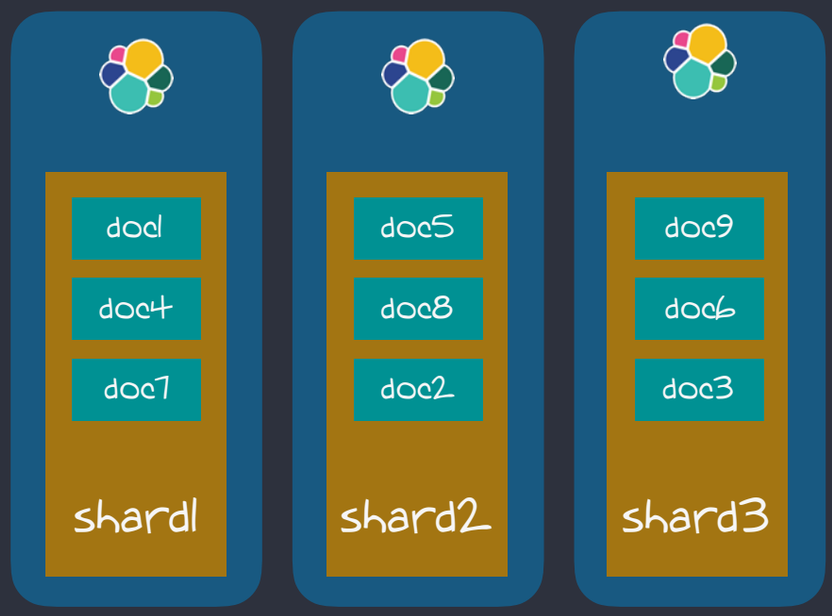
How It Works

ElasticSearch and Spark work great together.

Here is the idea. Say you have elastic search data distributed over a number of nodes. You want to run a Spark job on that data.

How do you do it ?

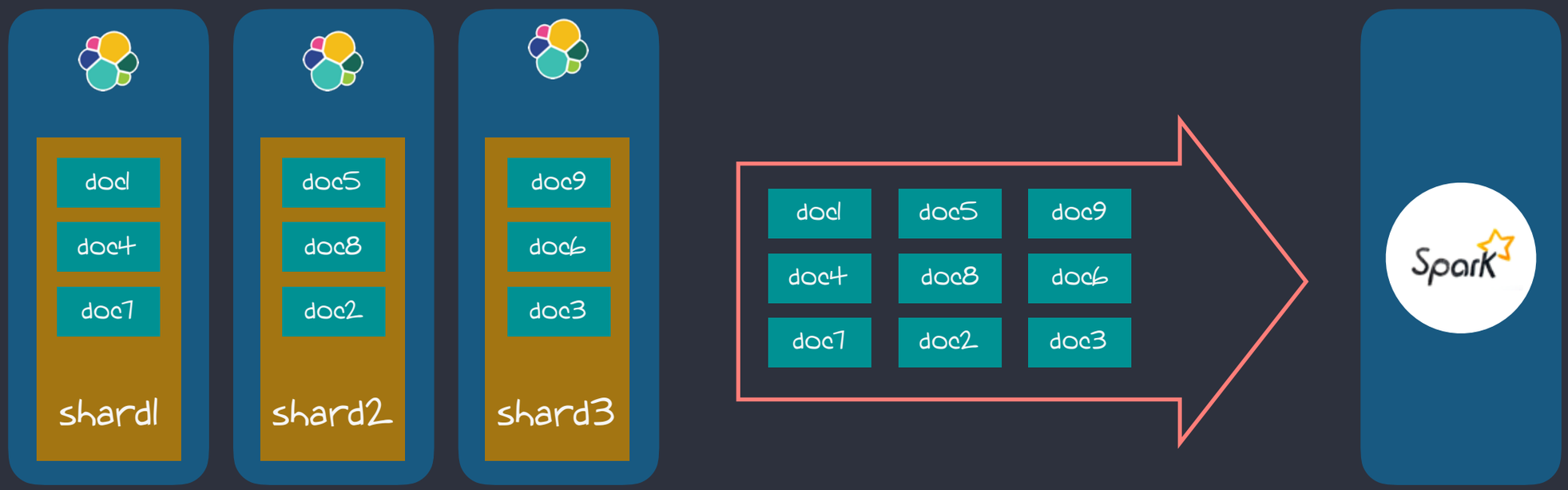
Checkout http://punchplatform.io for more information.



Doing It Wrong

Using a naive approach, you will run your Spark processing in a separate Spark cluster. That will result in reading all the required data from Elasticsearch and transfer it to Spark.

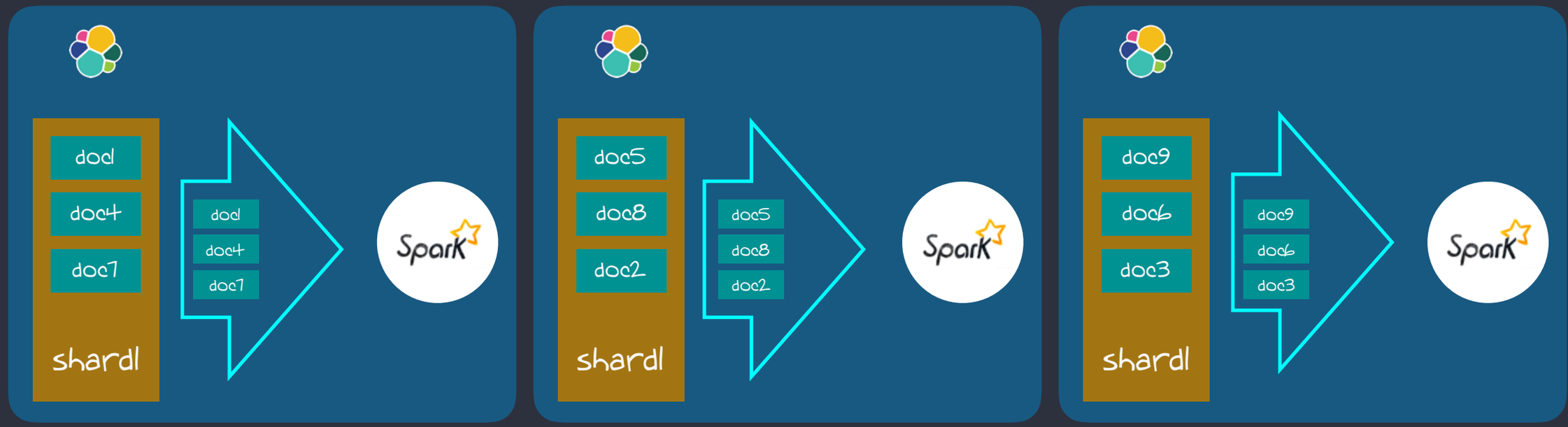
Not a good idea.



Doing It Right

Punchplatform lets you deploy Elasticsearch and Spark nodes on the same servers, equipped with the Elasticsearch spark-hadoop connector.

Every Job access the Elasticsearch node local data



Thanks !