



Machine Learning
the PunchPlatform way



PunchPlatform team



Agenda

- Spark Made Simple
- Fit Job
- Transform Job
- Leveraging Spark MLlib

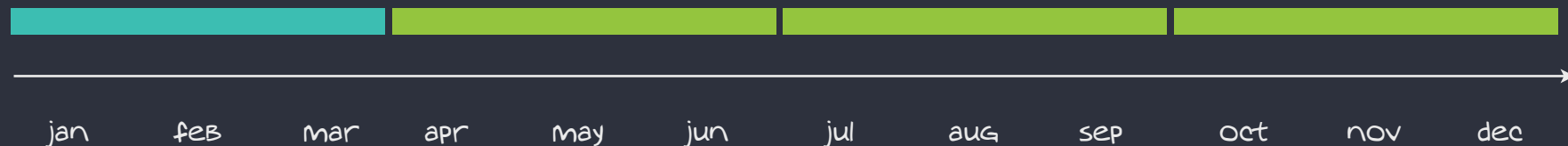


Problem Statement



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

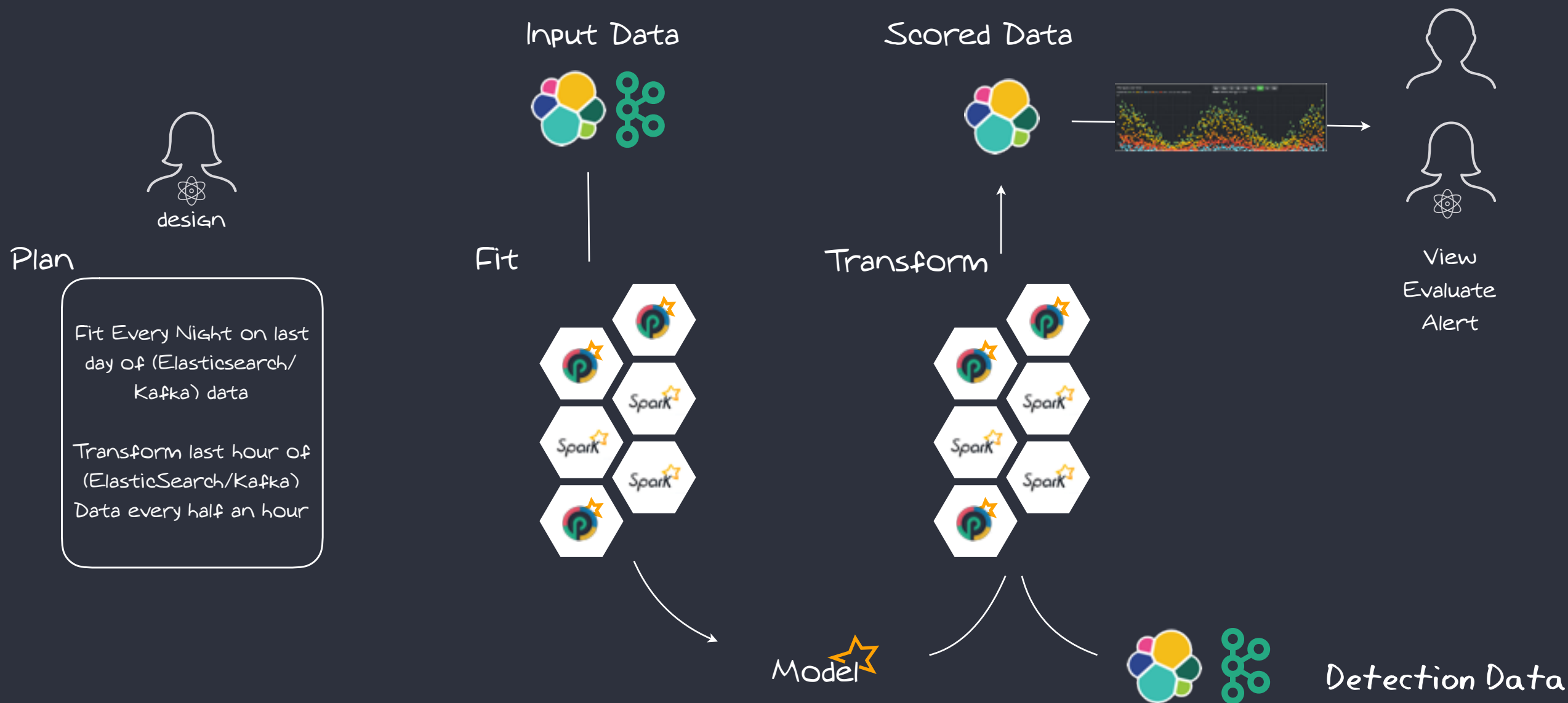
- **react** : Using Kibana and the power of Elasticsearch query language, you keep an eye to what is going on you system.
- **report** : compute powerful aggregated indicators overs last weeks of data. You keep control of important or suspect activities.
- **investigate** : perform on demand search and 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 ?



💡 Fit JOB Example



Punch Stage
Field selection, enrichment, ..

Spark Stage
K-Means, Regression, ...

Model Output
Save the computed Model



Data Input
ElasticSearch last day of
firewall logs

Spark Stage
SparkQL

Spark Stage
Filter

💡 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



<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,...



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.



Thanks !