© 2018 SPLUNK INC.

Using Spark and MLlib for Large Scale Machine Learning With Splunk Machine Learning Toolkit

Lin Ma, Principal Software Engineer
Fred Zhang, Principal Data Scientist

© 2018 SPLUNK INC.

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC.

The forward-looking statements made in this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

Splunk, Splunk>, Listen to Your Data, The Engine for Machine Data, Splunk Cloud, Splunk Light and SPL are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners. © 2018 Splunk Inc. All rights reserved.

© 2018 SPLUNK INC.

What is Machine Learning

- A process for generalizing from examples
- Examples

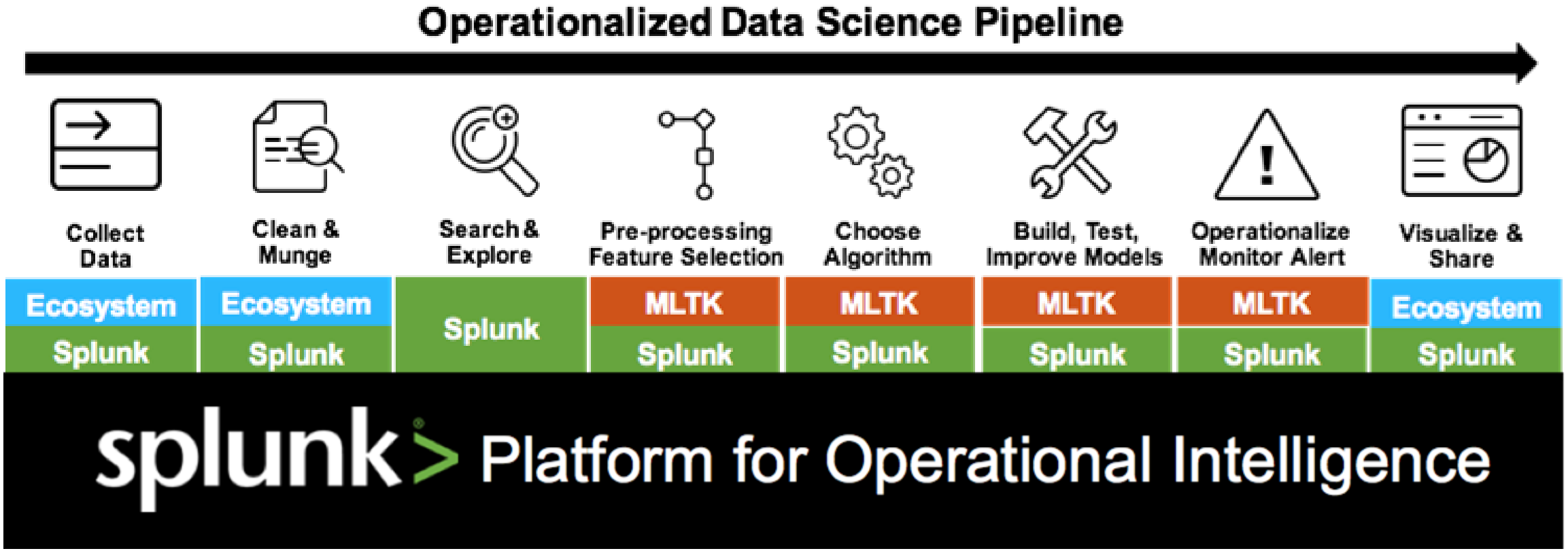
A, B, … → #	(regression)
A, B, ... → a	(classification)
$X_{past} \rightarrow X_{future}$	(forecasting)
like with like	(clustering)
$\lvert X_{predicted} - X_{actual} \rvert >> 0$	(anomaly detection)

© 2018 SPLUNK INC.

Overview of ML at Splunk

© 2018 SPLUNK INC.

Splunk Machine Learning Toolkit (MLTK)



© 2018 SPLUNK INC.

Splunk Machine Learning Toolkit (MLTK)

- **Assistants:** Guided model building, testing and deployment for common objectives
- **Showcases:** Interactive examples for typical IT, security, business and IoT use cases
- **Algorithms:** 30 standard algorithms (supervised & unsupervised)
- **ML Commands:** New SPL commands to fit, test and operationalize models
- **ML-SPL API:** Extensibility to easily import any algorithm (proprietary / open source)
- **Python for Scientific Computing Library:** Access to 300+ open source algorithms
- **Connector for Spark**: Support large scale model training

© 2018 SPLUNK INC.

MLTK is a Bridge to Python for Scientific Computing and Custom ML Algorithms

scikit learn

SM StatsModels
Statistics in Python

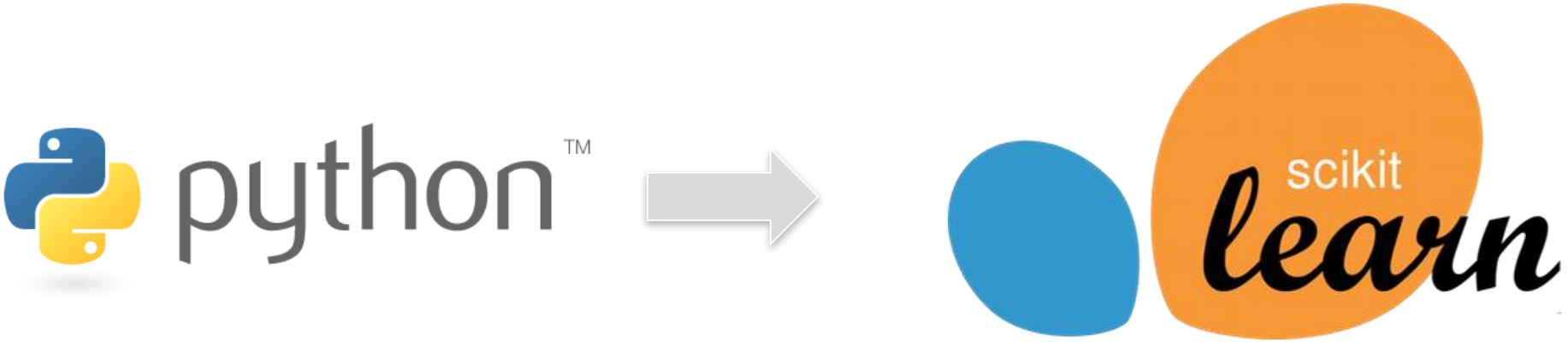
© 2018 SPLUNK INC.

What is Spark

- General-purpose data processing engine
- Optimized to run in memory
- Spark streaming and processing
- Interactive streaming analytics
- Data integration
- Machine learning

© 2018 SPLUNK INC.

What is Spark MLlib



© 2018 SPLUNK INC.

Splunk MLTK Connector for Apache Spark™

A bridge from MLTK to Spark MLlib

© 2018 SPLUNK INC.

Feature Highlights

- Large-scale ML training
  - Large data set from Splunk
  - Distributed/parallelized model training
  - Offload compute from Splunk search head

© 2018 SPLUNK INC.

Feature Highlights

- BYOSC (Bring Your Own Spark Cluster)
  - Utilize your existing compute resources
  - Use compute resources on demand
  - Minimal provisioning effort

© 2018 SPLUNK INC.

Feature Highlights

- Resource fine tuning
  - CPU cores
  - Memory limit
  - Return data volume tuning

© 2018 SPLUNK INC.

Feature Highlights

- Optimization to minimize data round-trip
  - New Syntax ***sparkml [sfit … | sfit …]***
  - Construct ML pipeline

splunk> .conf18

© 2018 SPLUNK INC.

What Problems are we NOT Solving (YET)

- NOT shipping Spark with splunk
- Only works with a predefined set of algorithms available in MLLib
- NOT designed to run arbitrary Scala code
- NOT designed to ingest data from outside of Splunk

© 2018 SPLUNK INC.

Supported Algorithm

- Regression
  - LinearRegression
  - DecisionTreeRegressor
  - RandomForestRegressor
  - GBTRegressor
- Preprocessing
  - PCA
- Classification
  - LogisticRegression
  - DecisionTreeClassifier
  - RandomForestClassifier
  - GBTClassifier
  - NaiveBayes
- Clustering
  - K-means

© 2018 SPLUNK INC.

Resource Managers

- Supported
  - Standalone
- Experimental
  - YARN
    - supports secure mode via kerberos
  - Mesos
- Coming Soon
  - Kubernetes

© 2018 SPLUNK INC.

When to use Connector for Apache Spark™

- The size of the training data is larger than the physical memory available on the search head
- Have a Spark cluster readily available
- Use the connector to take advantage of parallelizable algorithms in Spark MLLib

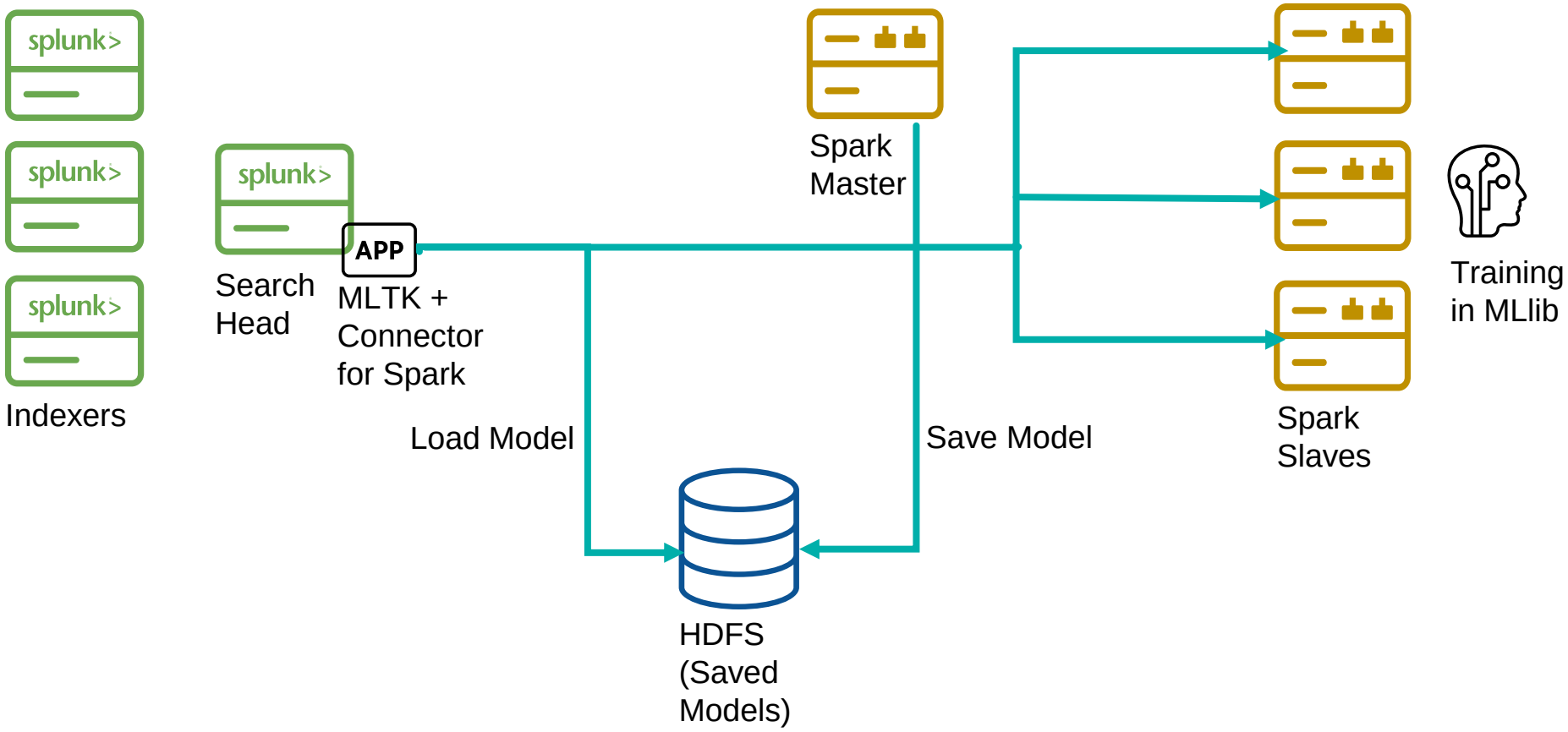
© 2018 SPLUNK INC.

Guidelines used for the Architecture

- Use existing technology and infrastructure
- Minimal configuration
- Easy resource management
- Efficiency

© 2018 SPLUNK INC.

Deployment Architecture



splunk> .conf18

© 2018 SPLUNK INC.

Requirements

- Splunk Enterprise 7.0+
- Splunk MLTK 2.3+
- Spark cluster 2.1+

© 2018 SPLUNK INC.

Connection Settings

sparkml.conf

Spark Standalone

Hadoop YARN

Mesos

```
[connection:example]
cluster_manager=standalone
spark.master=local[*]
search.search_head_address=<search head hostname>
```

```
[connection:example]
cluster_manager = yarn
spark.yarn.stagingDir = <location of staging dir on hdfs>
spark.hadoop.yarn.resourcemanager.hostname = <yarn resource
manager hostname>
```

© 2018 SPLUNK INC.

sparkml.conf

common configuration

Common Configurations

```
spark.executor.memory = 4g
spark.driver.memory = 4g
spark.driver.cores = 4
spark.cores.max = 24
spark.model_dir = ../spark-model/
```

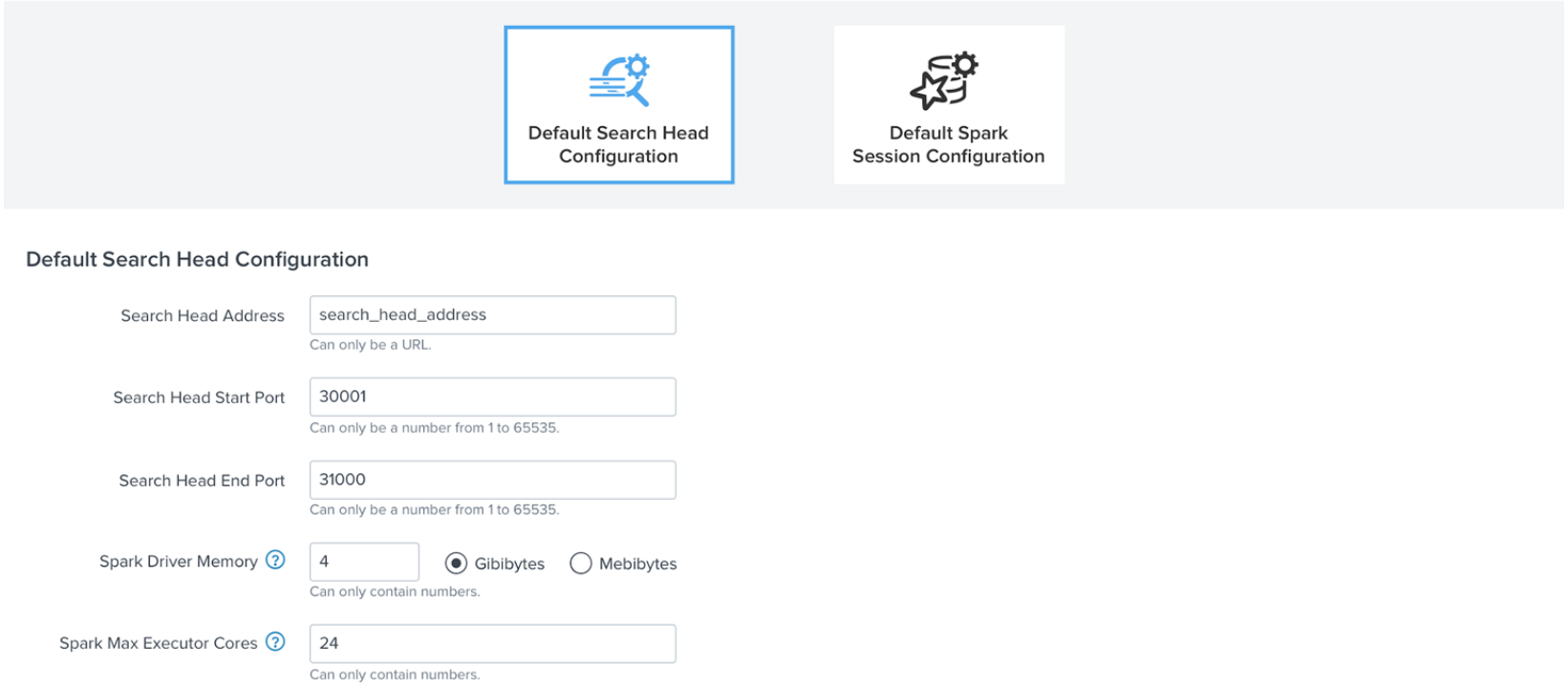
For a complete list of configurable options, refer to our documentations on : https://www.splunk.com/page/preview/mltkconnector

© 2018 SPLUNK INC.

UI Configuration

Hadoop YARN

experimental



© 2018 SPLUNK INC.

SPL Syntax

sfit

Syntax

```
sfit <algorithm>
        (<option_name>=<option_value>)*
        (<algorithm-arg>)+
        (into <model_name>)?
        (as <output_field>)?
```

Example

```
sparkml [
        sfit LinearRegression errors from _time into errors_over_time
        ]
```

© 2018 SPLUNK INC.

SPL Syntax

sapply

Syntax

```
sapply <model_name>
    (as <output_field>)?
```

Example

```
sparkml [
    sapply errors_over_time as predicted_errors
    ]
```

© 2018 SPLUNK INC.

SPL Syntax

sparkml

Syntax

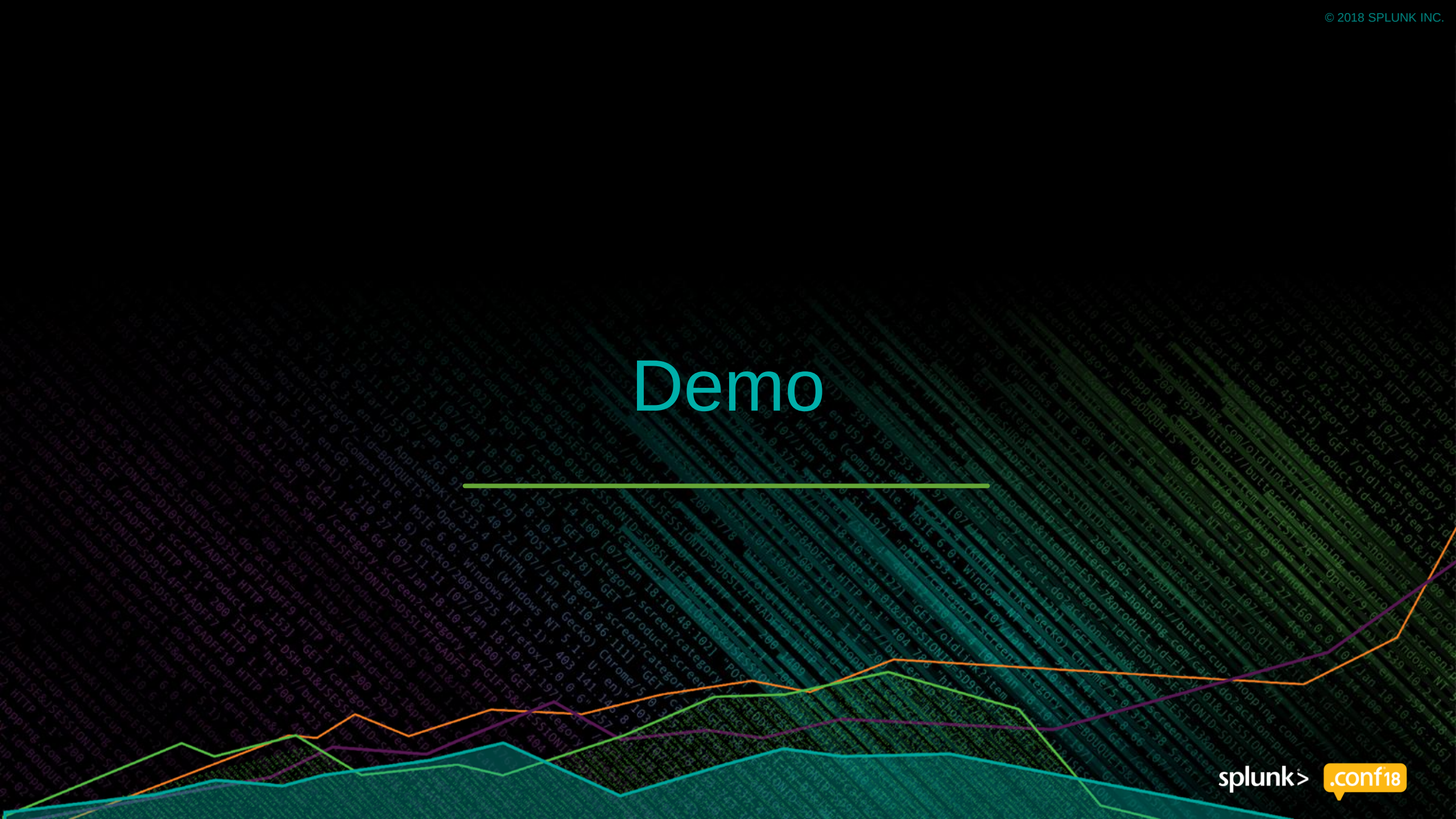
```
sparkml
        (<option_name>=<option_value>)*
        [ <sfit ...> | <sapply ...>*?]
```

Example

```
sparkml [
        sfit PCA k=3 from field1 field2 field3 as pca_output |
        sfit LinearRegression errors from _time pca_output_1
                pca_output_2 pca_output_3
        ]
```

© 2018 SPLUNK INC.

Demo



© 2017 SPLUNK INC.

Splunk MLTK Connector for Apache Spark™

Version 0.9.0

splunk> listen to your data

© 2018 SPLUNK INC.

Contact Us

- Beta program feedbacks: sparkml@splunk.com
- We want to hear from you if the add-on provides value ☺
- We also want to hear about things we need to improve
  - Any bugs you run into
  - Failed use cases
    - Why did it fail
    - Why is it important to your business
    - What we can do to make it successful
  - Additional features you want
    - functional: such as new algorithm
    - non-functional: such as more throughput
- Feedback is critical for us to deliver products that bring success to our customers

© 2018 SPLUNK INC.

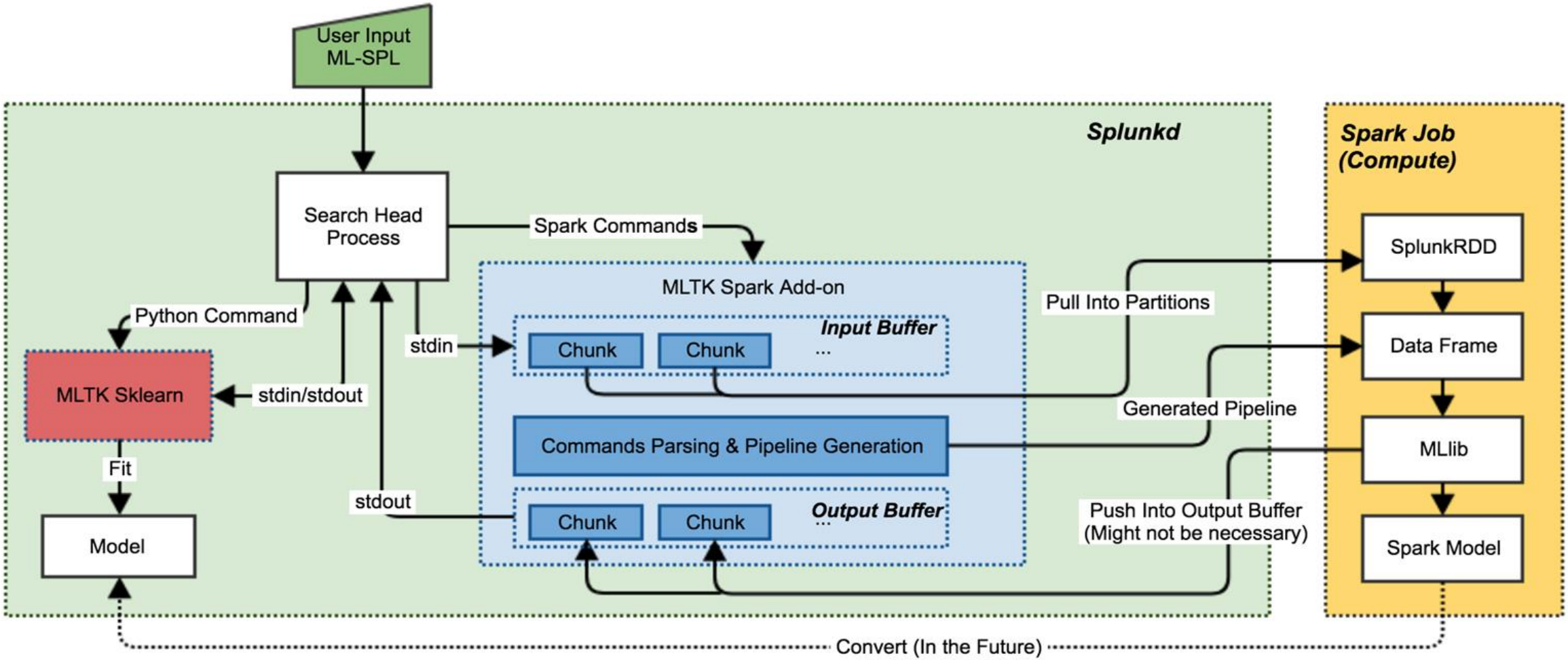
Thank you

Don't forget to rate this session in the .conf18 mobile app

© 2018 SPLUNK INC.

Appendix

Detailed Architecture



splunk> .conf18