DATABIZ
your solution builder

Going Big Data on Apache Spark

KNIME Italy Meetup

Agenda

Introduction
Why Apache Spark?

Section 1
Gathering Requirements

Section 2
Tool Choice

Section 3
Architecture

Section 4
Devising New Nodes

Section 5
Conclusion

Let’s dive

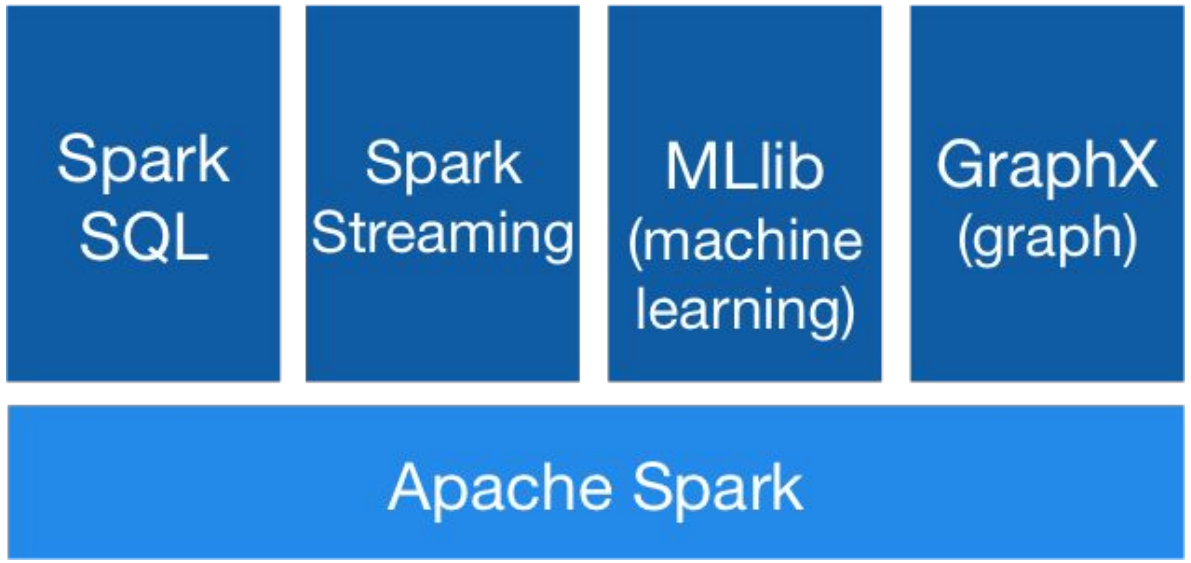
DATABIZ
your solution builder

Introduction

Why Apache Spark?

Apache Spark

- Fast engine for large-scale distributed data processing
- Builds and tests predictive models in little time
- Built-in modules for:
  - SQL
  - Streaming
  - Machine Learning
  - Graph Processing



(Source: Apache Spark)

Dealing with Apache Spark

Apache Spark has a steep learning curve

Use a data scientist-friendly interface with Spark integration

Section 1

Gathering Requirements

Requirements

Tasks	Required Capabilities
Explore data on Hadoop Ecosystem	Hadoop integration
Leverage Spark for fast analysis	Spark integration
Data Preparation	Modeling
Perform statistical analysis	Extensibility

DATABIZ

Section 2

Tool Choice

(Source: Gartner, February 2015)

Product	Suitable	Notes
Alpine	✓	
KNIME	✓	
RapidMiner	✓	
KXEN	?	Spark integration on the roadmap
SAS	?	Spark integration on the roadmap
IBM SPSS	?	Spark integration on the roadmap

What About KNIME?

- User-friendly interface
- Open Source
- Integration with Hadoop / Spark
- Clear pricing and cost effectiveness
- Rich existing feature set
- Possibility of co-development

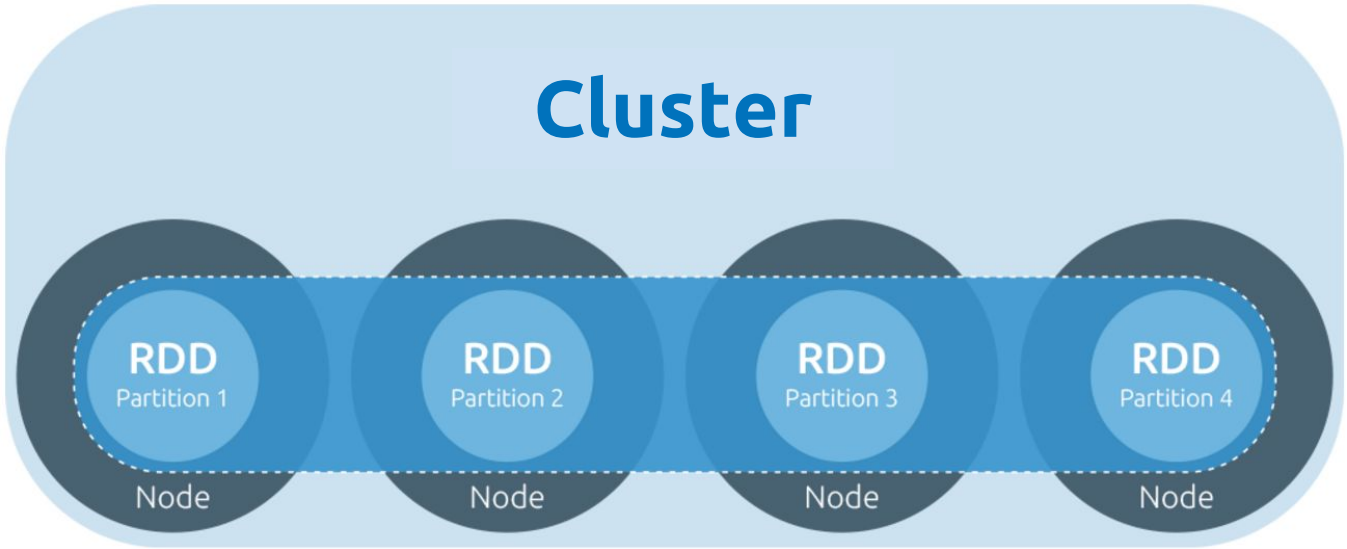
DATABIZ
your solution builder

(Source: Gartner, February 2015)

Section 3

Architecture

Spark's Building Block: RDD



Immutable

Each step of a dataflow will create a new RDD

Lazy

Data are processed only when results are requested

Resilient

A lost partition is reconstructed from its lineage

From KNIME to Spark: Spark Job Server

- Simple REST interface for all aspects of Spark (job and context) management
- Support for Spark SQL, Hive and custom job contexts
- Named RDDs and DataFrames: computed RDDs and DataFrames can be cached with a given name and retrieved later on

Spark Job Creation

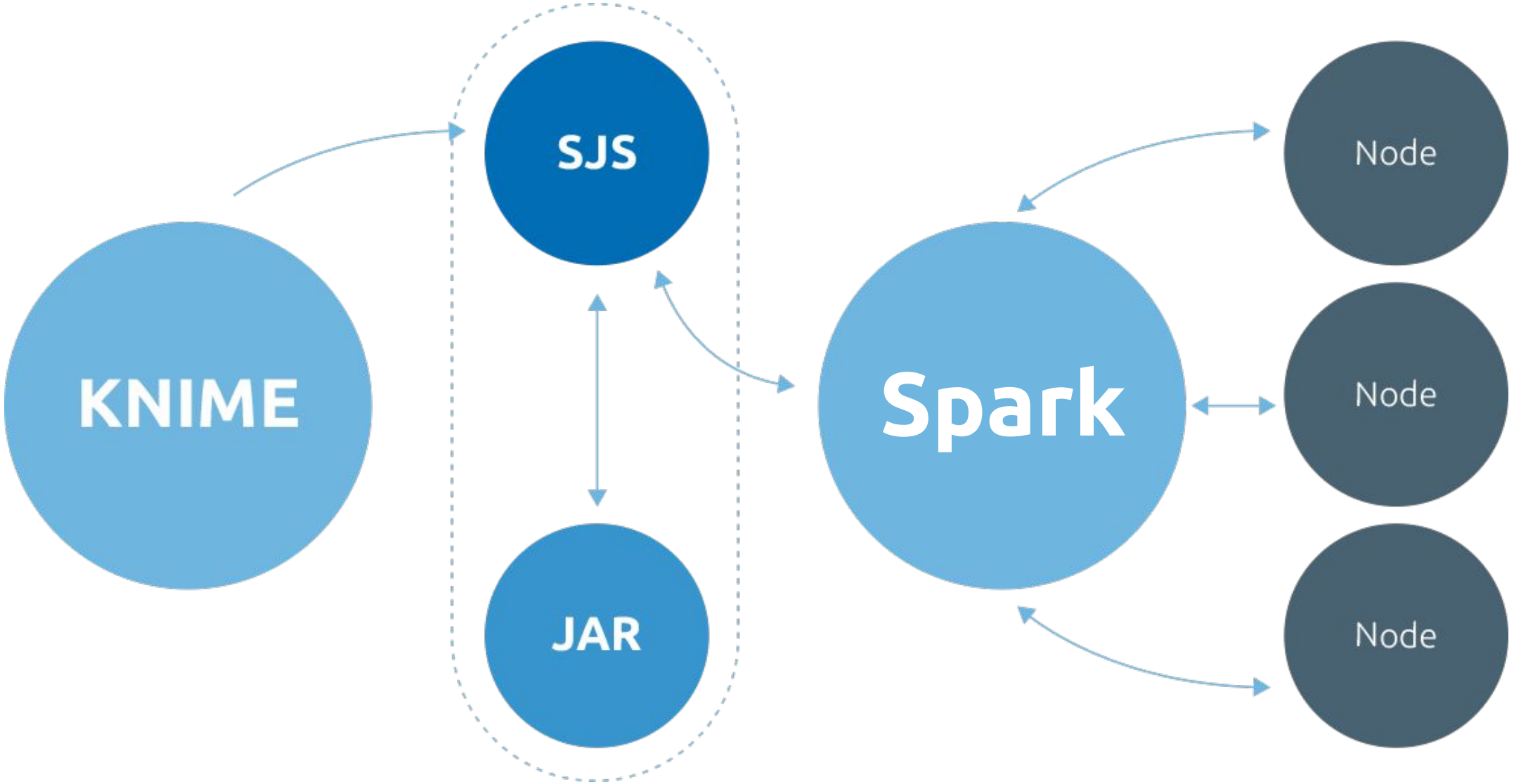
To create a job that can be submitted through the job server, the job must extend the **SparkJob** trait. Your job will look like this:

```
object SampleJob extends SparkJob {

    override def validate(sc: SparkContext, config: Config): SparkJobValidation = ???

    override def runJob(sc: SparkContext, jobConfig: Config): Any = ???

}
```

Spark Job Structure

validate allows for an initial validation of the context and any provided configuration. It helps you preventing running jobs that will eventually fail due to missing or wrong configuration and save both time and resources.

runJob contains the implementation of the Job. The SparkContext is managed by the JobServer and will be provided to the job through this method.


KNIME
SJS
JAR
Spark
Node
Node
Node

Warning!

Spark Job Server won't send updates to KNIME regarding its internal state

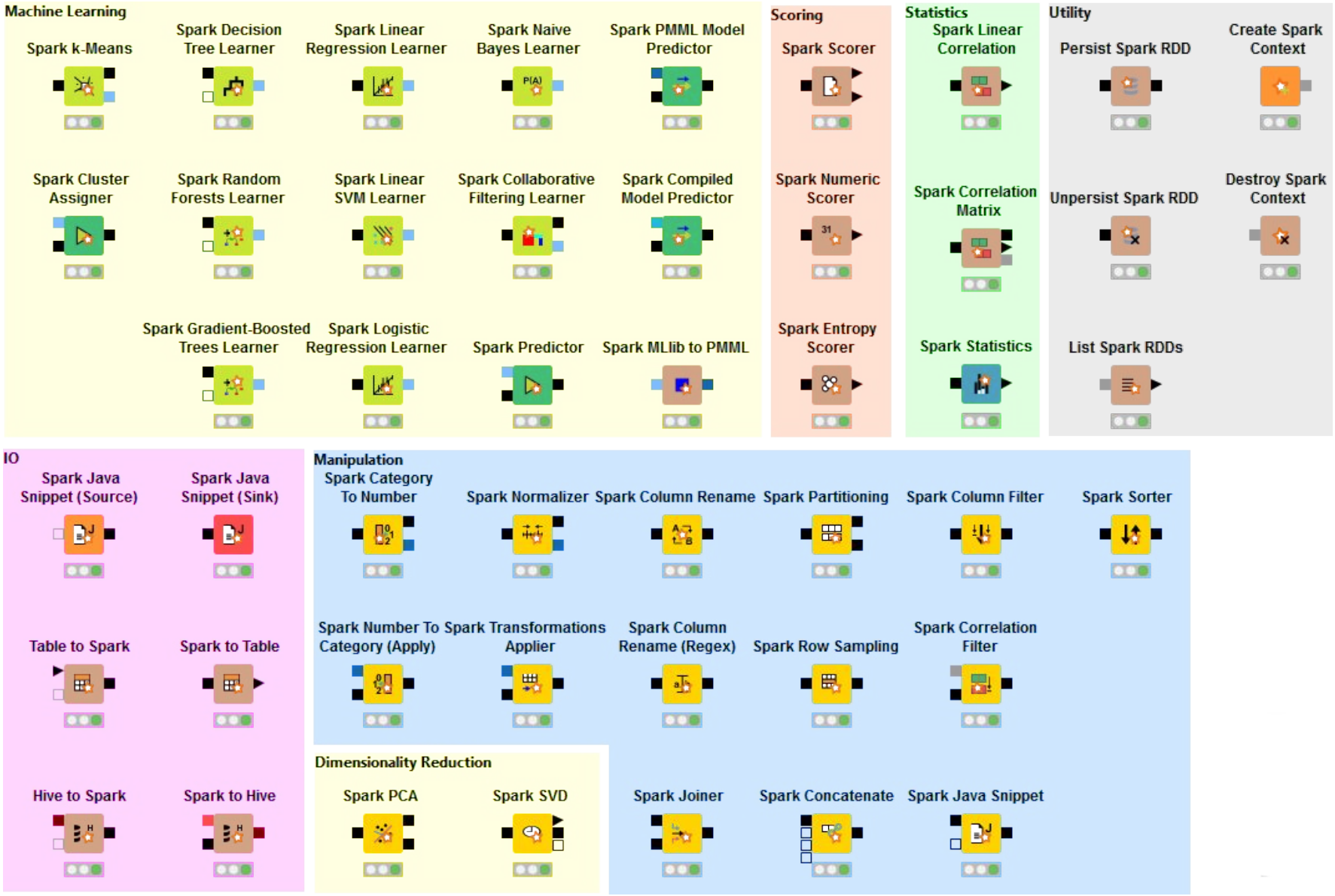
This means that, if Spark Job Server is down or is restarted, you will never know it!

Due to lazy evaluation, functions are computed only when the result is actually required

This means that you can't always trust KNIME's green light!

Section 4

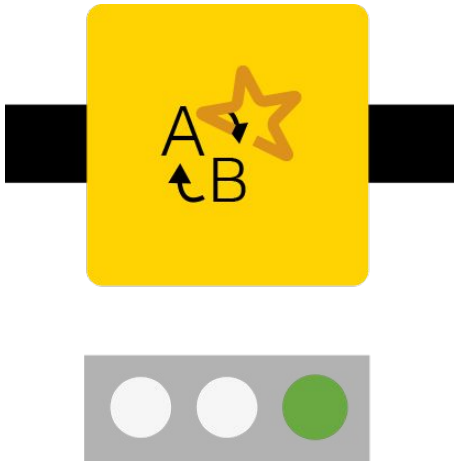
Devising New Nodes



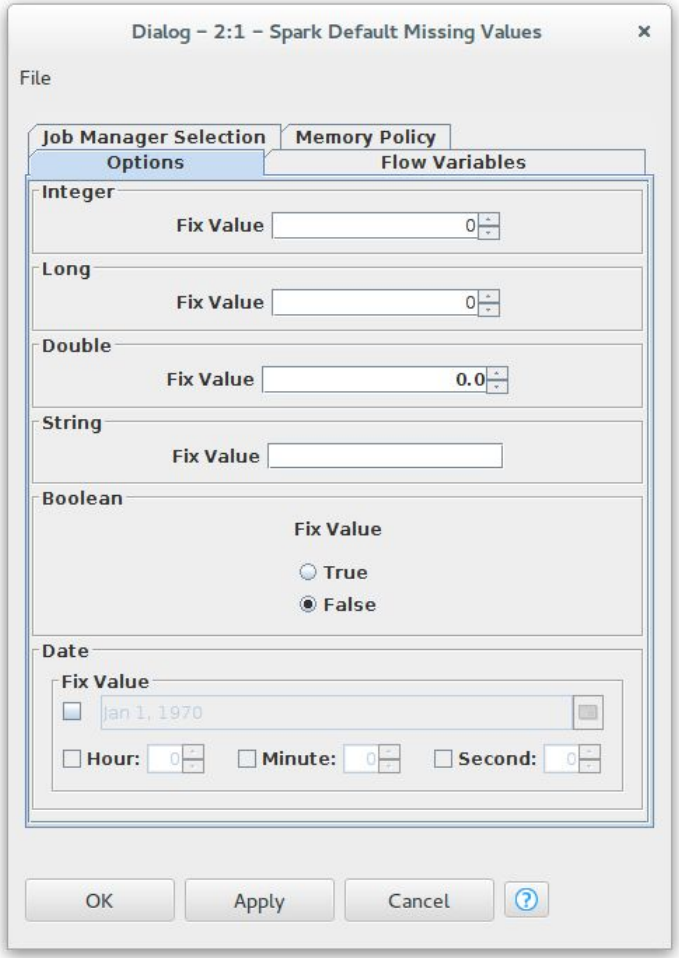
(Source: KNIME)


I WANT MORE
MORE! MOAR!! MOOOAAARRR!!!
quickmeme.com

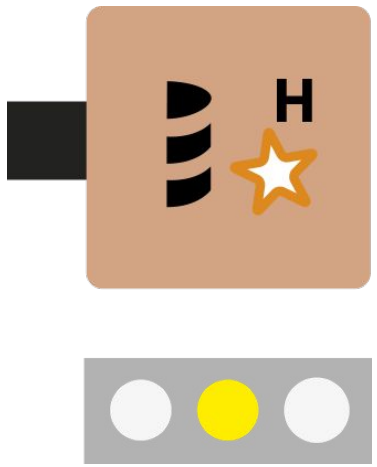
Spark Default Missing Values



- Replaces all occurrences of missing values with fixed default values.
- Can be applied to:
  - Integers
  - Longs
  - Doubles
  - Strings
  - Booleans
  - Dates

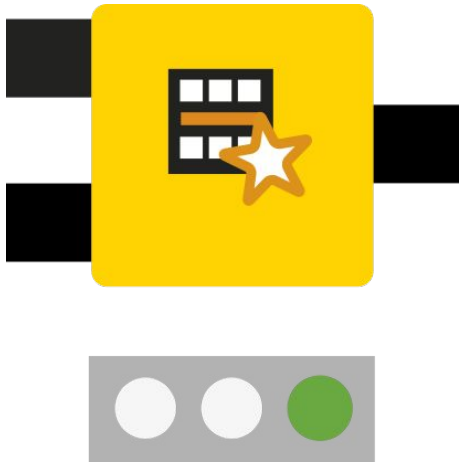


Spark to HBase

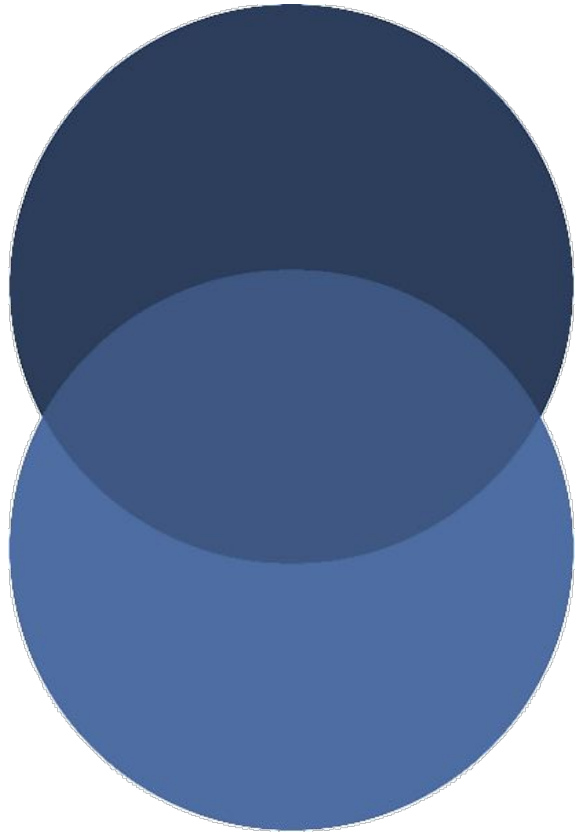


- Persists data from an incoming RDD into HBase.
- Requires:
  - the name of the table you intend to create
  - the name of the column in which the Row IDs are contained
  - the name that will be given to the Column Family

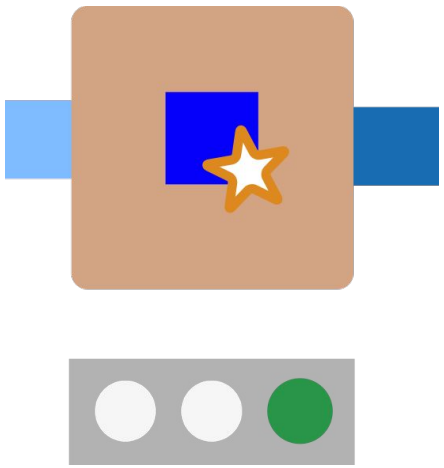
Spark Subtract



- Given two incoming RDDs, returns an element from the first input port RDD that are not contained in the second input port RDD.
- Similar to the relative complement of the set theory



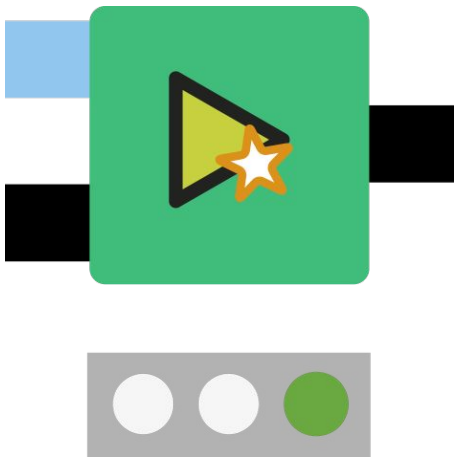
Spark MLlib to PMML



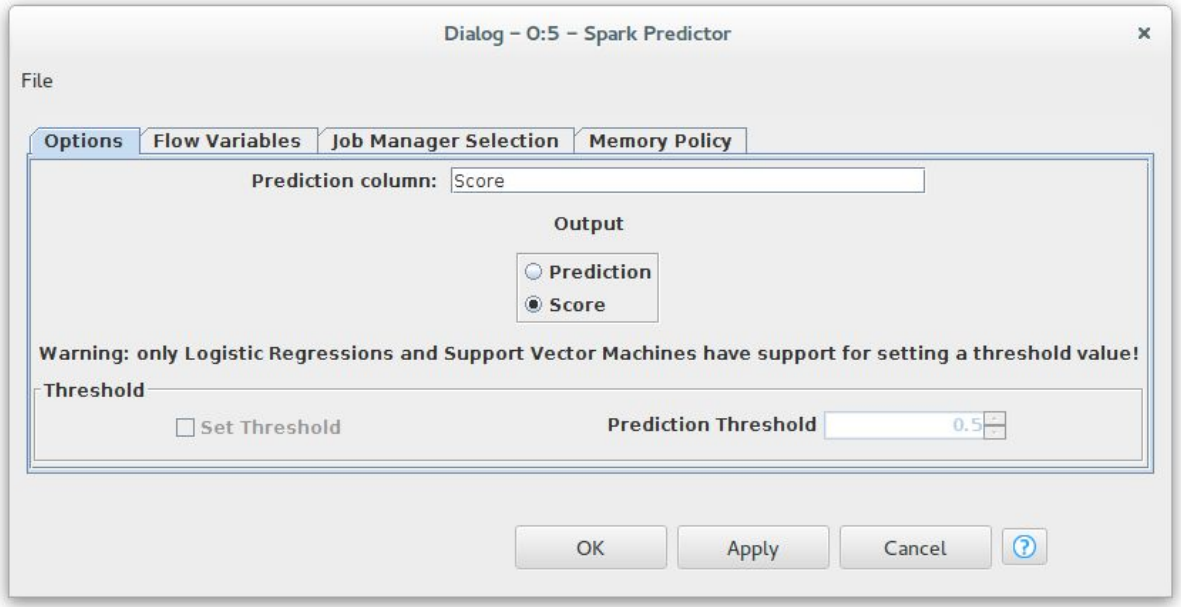
- Converts supported Spark MLlib models into PMML files.

- Supported model are (up to now):
  - Decision Tree
  - K-Means
  - Linear Regression
  - Logistic Regression
  - Naive Bayes
  - Random Forest
  - Support Vector Machines

Spark Predictor + Scoring

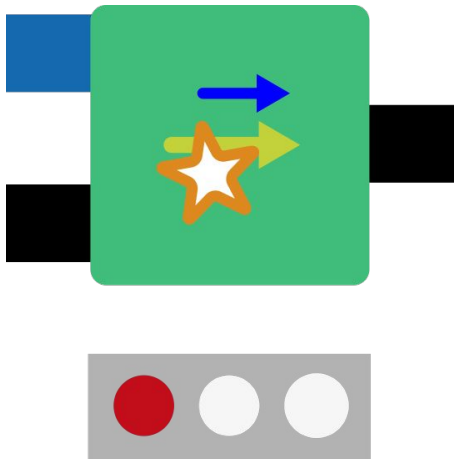


- Labels new data points using a learned Spark MLlib model
- Allows to get the score (the probability that an observation belongs to the positive class)
- Allows to set a threshold



Interlude:
Spark JPMML Scorer

Once Upon a Time: Spark PMML Predictor



"Compiles the given PMML model into bytecode and runs it on the input data on Apache Spark"

Spark PMML Predictor: Issues

- Returns a prediction, not a score

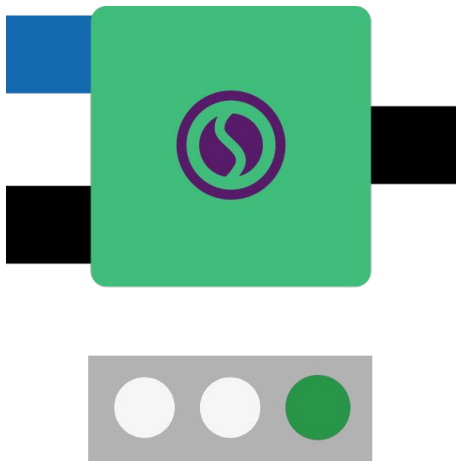
- *"Compiles the given PMML model into bytecode"*
  - not testable
  - java.lang.ClassFormatError: Invalid method Code length
    - maximum code length is 65534 bytes!

Solution: JPMML

“Java API for Predictive Model Markup Language (PMML)”

- Adopted by the community
- Thoroughly tested (and testable!)
- Easily customizable
- Well documented

Spark JPMML Model Scorer



- Sends the PMML file to the Apache Spark cluster
- Takes advantage of the JPMML library to turn the PMML file into a predictive model
- Uses the JPMML library to give a score to the model's prediction

Other Activities

BugFixing: Spark to Hive & Hive to Spark

- Scala, Java and Hive all have different types
- E.g.:
  - Scala: Int
  - Java: int / Integer
  - Hive: INT
- All of these conversions were handled

Support: Database Connector + Impala

*jdbc:impala://your.host:1234/;auth=noSasl;**UseNativeQuery=1***

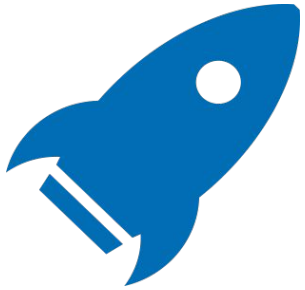
- Enabling the UseNativeQuery option, no transformation is needed to convert the queries into Impala SQL
  - If the application is Impala aware and already emits Impala SQL, enabling this feature avoids the extra overhead of query transformation
- Moreover, we noticed that this solves concurrency issues in Impala

DATABIZ
your solution builder

Section 5

Conclusion

Conclusion



- Apache Spark can run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk

- KNIME gives a data scientist-friendly interface to Apache Spark

- Yet, when dealing with Spark, even from KNIME, a (basic!) understanding of its inner workings is required

Thanks!

Special Thanks

- Sara Aceto
- Stefano Baghino
- Emanuele Bezzi
- Nicola Breda
- Laura Fogliatto
- Simone Grandi
- Tobias Koetter
- Gaetano Mauro
- Thorsten Meinl
- Fabio Oberto
- Luigi Pomante
- Simone Robutti
- Stefano Rocco
- Riccardo Sakakini
- Enrico Scopelliti
- Rosaria Silipo
- Lorenzo Sommaruga
- Marco Tosini
- Marco Veronese
- Giuseppe Zavattoni

DATABIZ


Hey,
We Are
Hiring!
Send Us Your CV!
hr@databiz.it
DATABIZ
your solution builder

Hey, We Are Hiring!

Send Us Your CV!

hr@databiz.it

DATABIZ