FIAP

MBA+

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

MBA+

Plataformas Cognitivas

Prof.: MARCIO JUNIOR VIEIRA
Email: marcio@ambientelivre.com.br

Pentaho Data Mining - Weka

Weka

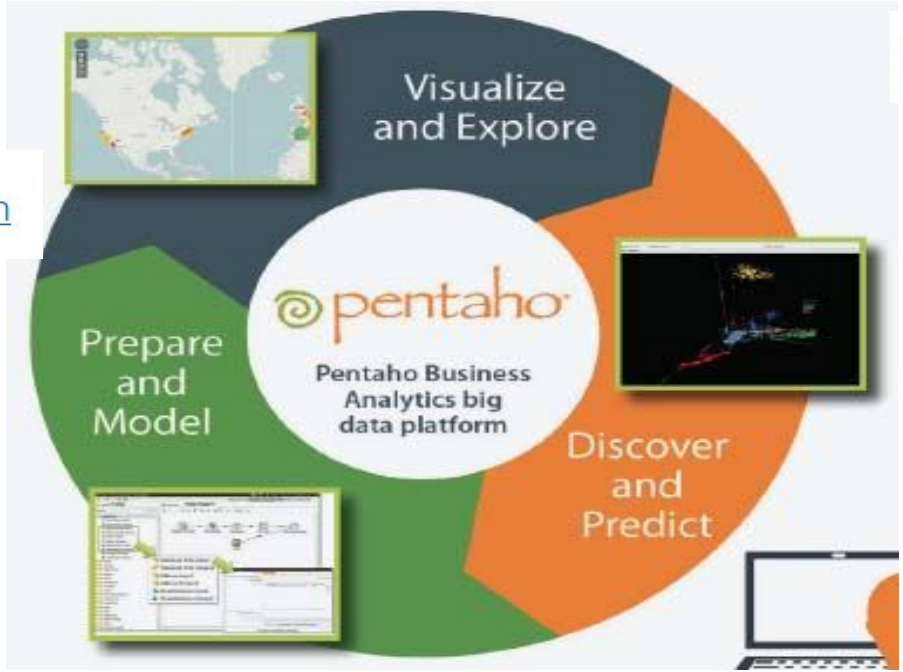
- Desenvolvido pela Universidade de Waikato (**W**aikato **E**nvironment for **K**nowledge **A**nalysis)
- Licença GPL
- Desenvolvido em Java
- Iniciado o desenvolvimento em 1993.
- O software foi adquirido pela Pentaho Corporation em 2016 (Hoje chamada de Hitachi Vantara).

- Site do projeto: http://www.cs.waikato.ac.nz/ml/weka/

Pentaho - Hitachi Vantara

- Plataforma abrangente para integração de dados e Business Analytics. 3 Pilares do Pentaho

Data Integration



Business Analytics

Big Data

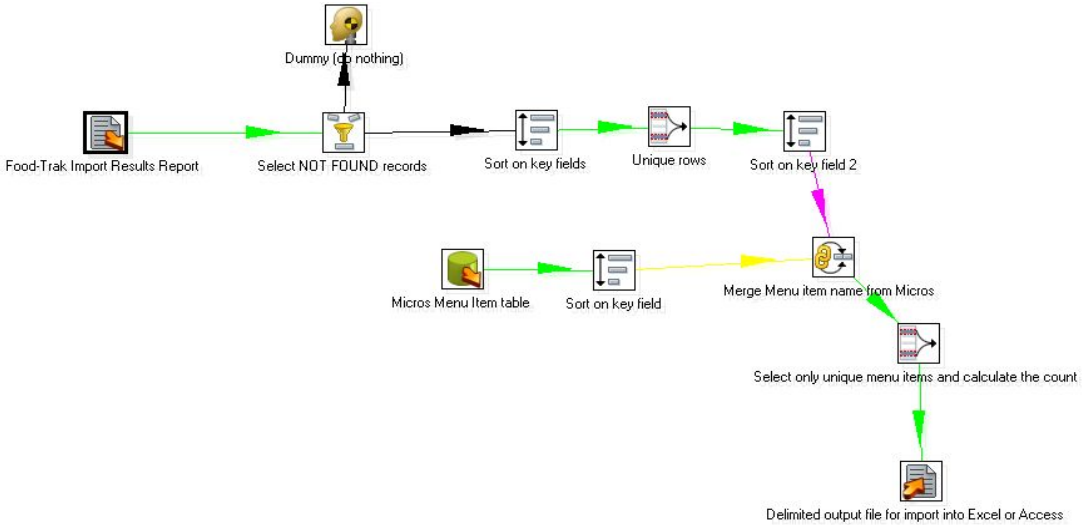
Pentaho Data Integration - PDI

Data Integration

- Pentaho Data Integration (PDI, também chamado Kettle) é um componente da suíte do Pentaho responsável pelos processos de Extração, Transformação e Carga (ETL).

ETL

- ETL - Extract, transform, load.



kettle™
pentaho data integration

THE DEVELOPER'S CONFERENCE

Pentaho Data Integration - Funcionalidades

Funcionalidades Tradicionais

- Usadas em projetos de data warehouse

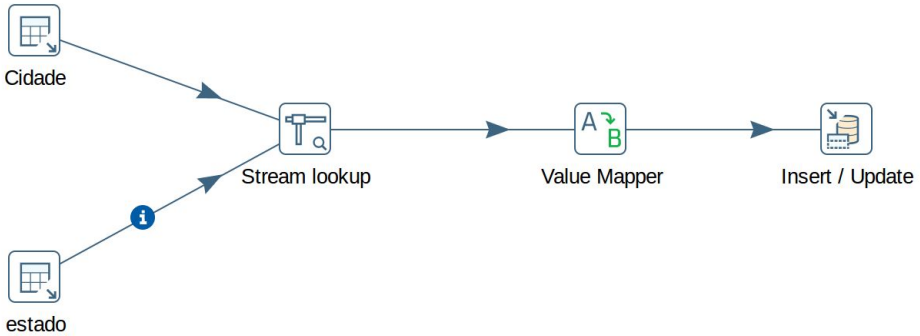
Funcionalidades Adicionais

- Migração de dados entre aplicações/banco de dados
- Exportar dados de banco de dados para arquivos texto
- Carregar massivamente dados em banco de dados
- Data Cleansing – disciplina de qualidade/limpeza de dados de data warehouse
- Integração de aplicações.
- Gerenciamento de Filesystem (File management)

O que é uma Transformação?

Definição

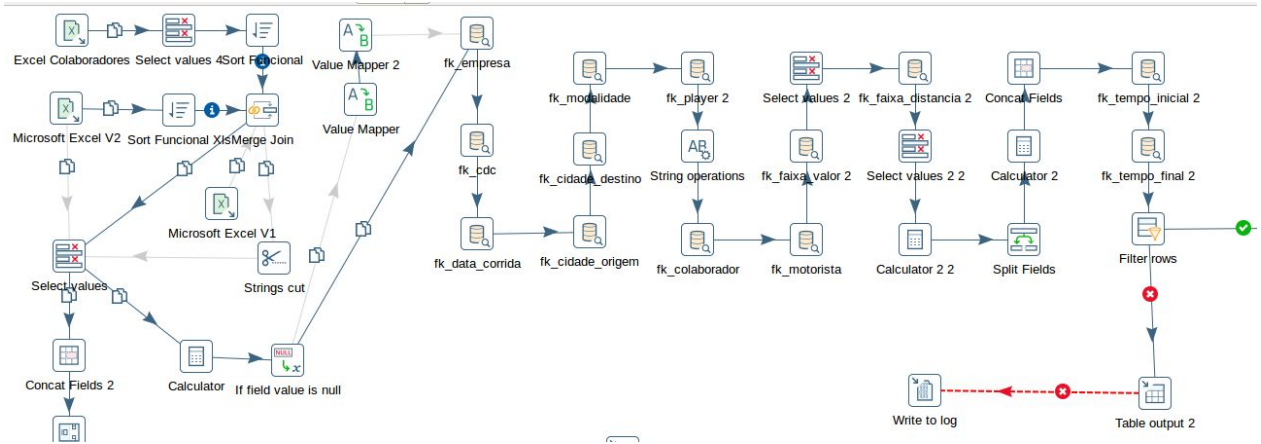
- Rotina com coleção de passos interligados
- Os primeiros são a fonte de dados.
- O último representa a saída de dados.
- Pode ser colocadas varias fontes de dados e saídas ou entrada
- É recomendado 1 transformação para cada dimensão ou tabela fato



O que são Steps?

Definição de Steps

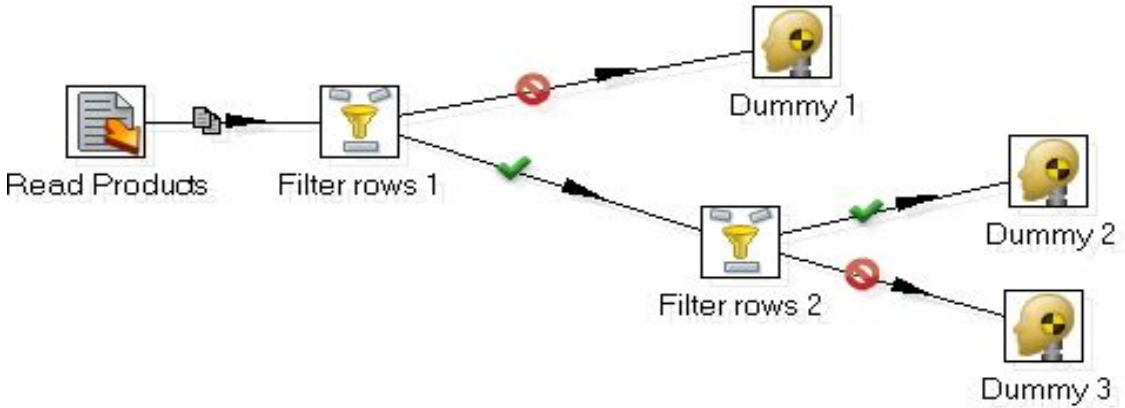
- Um passo é uma unidade mínima dentro de uma transformação.
- Grande variedade de passos
- Agrupada em categorias (input , Output, etc)
- Os tipos básicos são :
  Entrada, Transformação, Saída.



O que são Hops?

Definição

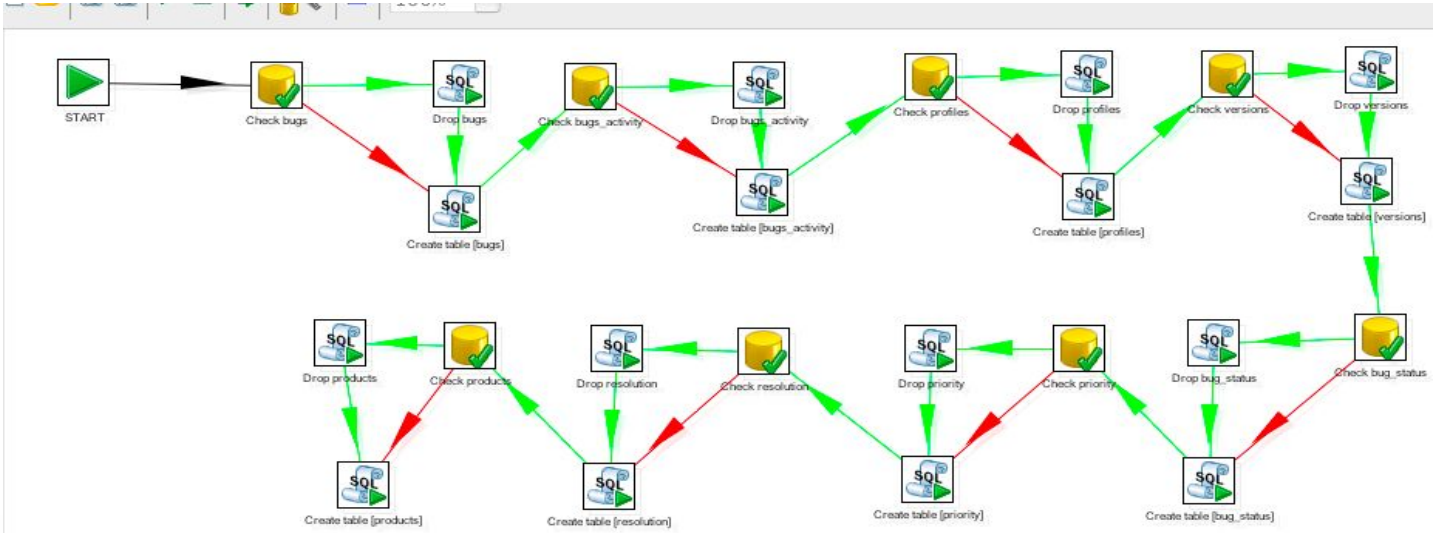
- Representação gráfica do fluxo de dados entre dois passos (conexão)
- Um deles Origem e outro Destino.



O que são Jobs?

Definição

- É uma rotina de execução
- Pode executar uma ou mais transformações
- Utilizado para cargas de tabelas fatos



A História de um Pedro!

Pedro

- Pedro queria trabalhar com TI mas ainda não tem uma especialidade.
- Pedro “ama dados”
- Pedro é muito estudioso!
- Pedro é “brasileiro e não desiste nunca”!!!
- Pedro leu que Data Scientist e um dos cargos mais “Sexys do Mundo”

A História do Matt!

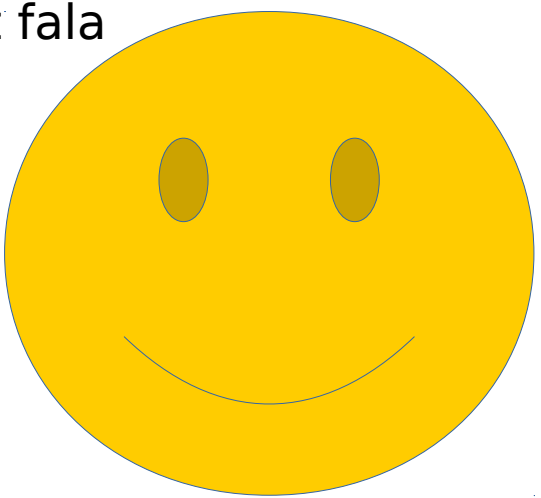
Matt

- Matt trabalha com dados a 20 anos
- Matt desenvolve open sources e software livres.
- Matt tem diversos apoiadores em seus projetos que colaboram com ideia, revisões, documentações, e melhorias.

Pedro e Matt!

Pedro

- Pedro conheceu matt de um blog
- Pedro se inspira em Matt e tenta aprender sobre tudo que Matt fala em seu blog...
- Mas Matt não está sozinho...

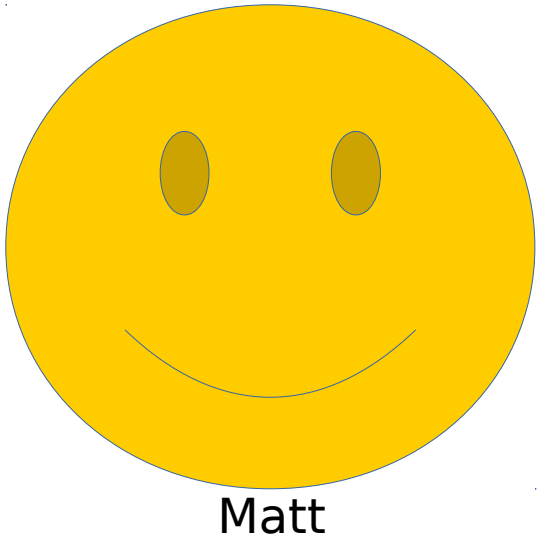


Matt

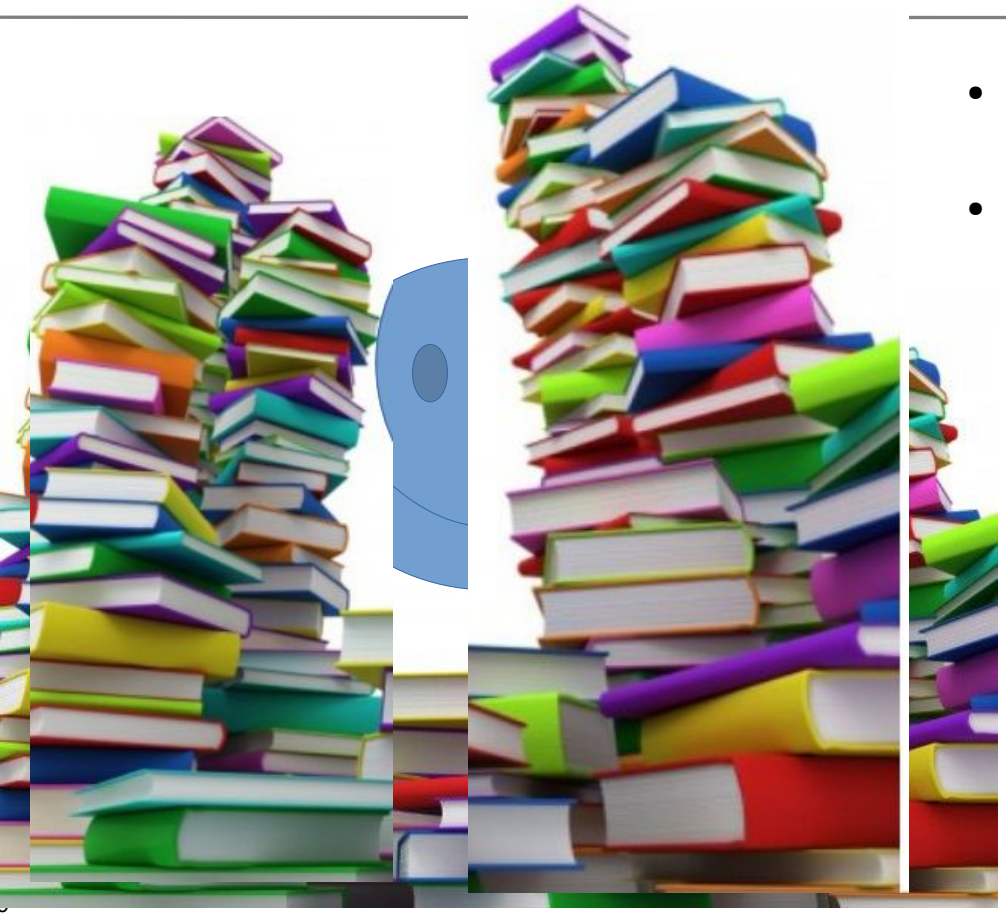
Pedro e Matt!

- Pedro compra tudo sobre Java Mapreduce e Hadoop e começa a estudar

- Matt está usando Hadoop MapReduce e Programando em Java, Pig, manipulando com Hive



Pedro e Matt!



- Hadoop Mapreduce 100% mais lendo que Spark Java verboso...
- Agora e Spark e Scala!!! diz Matt

Matt

Pedro e Matt!

- Streaming, Real Time, Kafka, Nifi, Cloud, ORC, parquet, Apex, Flume, Knox, Tez, Deap Learning..
- Matt esta postando muito muito mesmo em 2016...

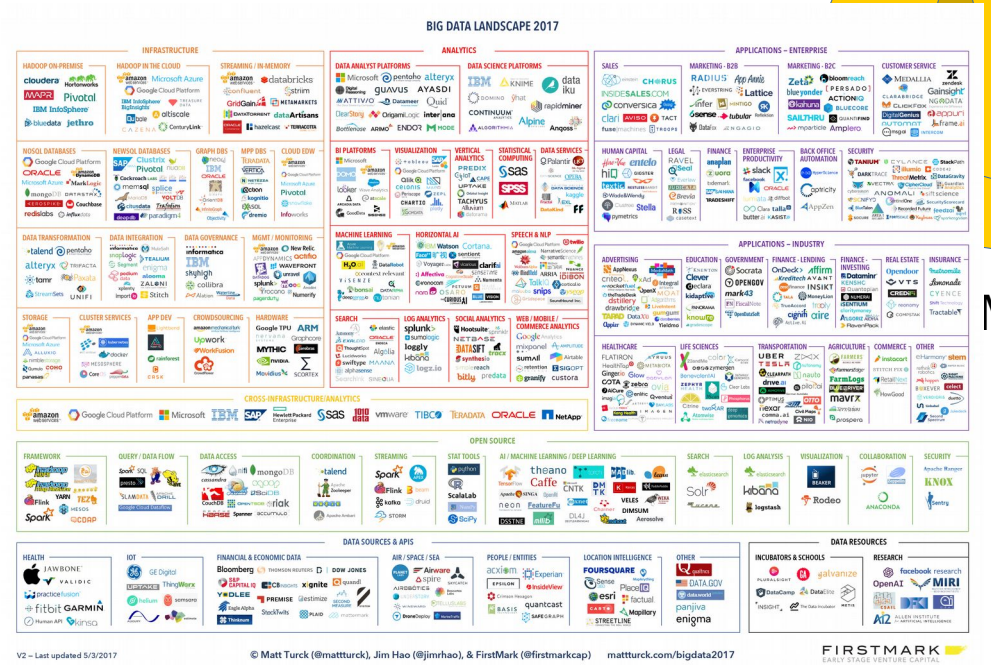
Matt

Pedro e Matt!

- Pedro entra na Faculdade de Administração!
- Biblioteca da cidade do Pedro recebe a maior doação de Livros da sua história

Pedro

- Matt apresenta a stack de Big Data, Data Science 2017



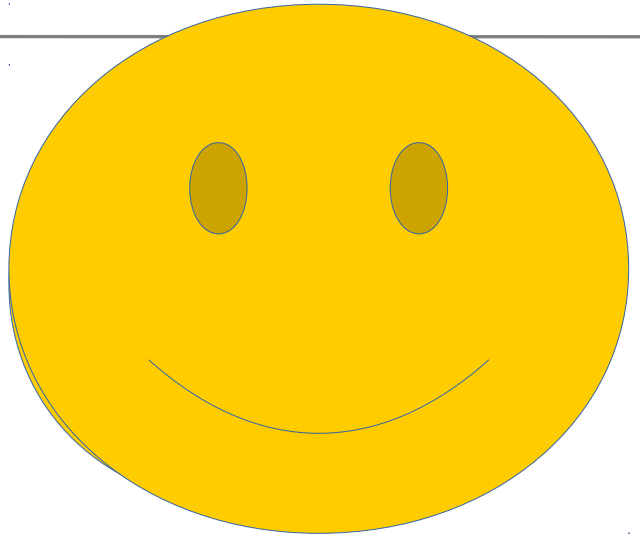
Matt

Pedro e Matt!

- Matt perde seu mais fiel seguidor
- E não deu tempo de avisar que Spark já não e mais tendência e agora e **Apache Flink** e ele poderia reusar o conhecimento de Java e Scala...

- E surge uma ideia!

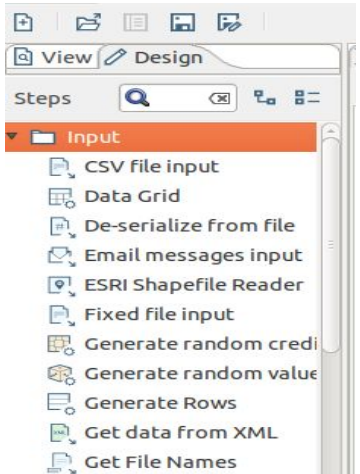
- Porque não **encapsular** seus conhecimento em um “framework” de trabalho que caso a tecnologia mude possamos manter os mesmos processo de trabalho com uma curva baixa de implementação e aprendizado, **alias o que o Pedro gosta e dos dados**!

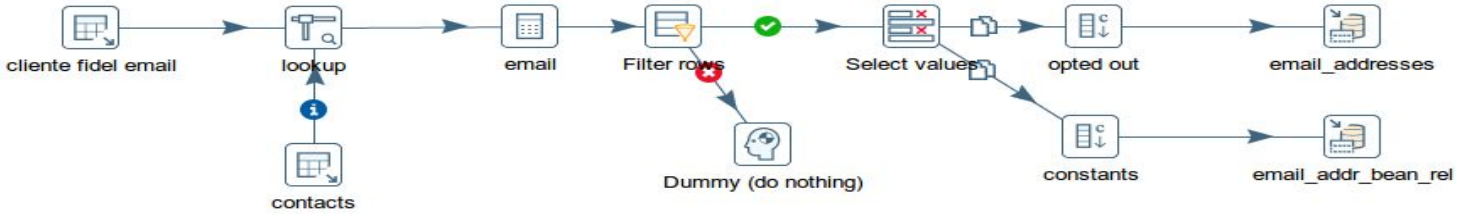


Matt

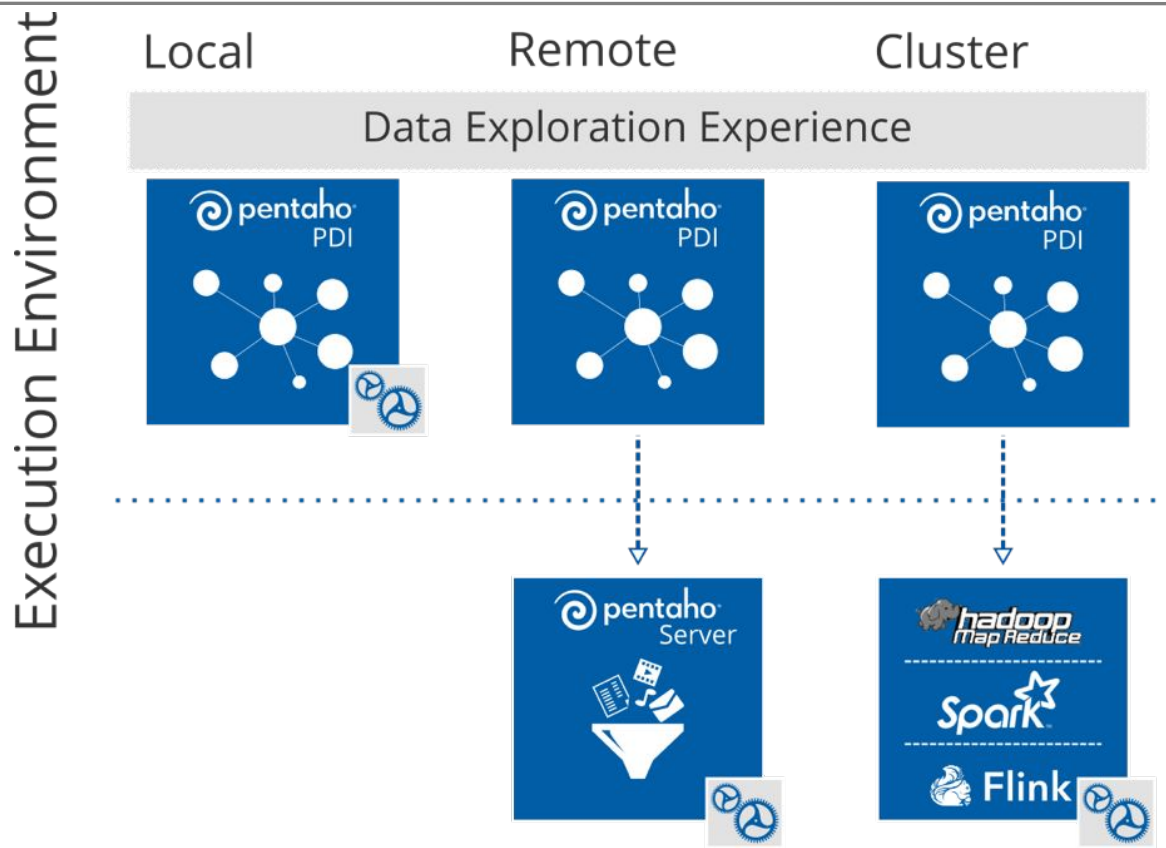
Pentaho Data Integration - PDI

- Processa em Paralelo Cluster Apache Spark
- Acessar dados diretamente (DW opcional)
- Permite publicar dados diretamente em Reports, Ad-Hoc Reports e Dashboards com uso integrado do Pentaho Server.
- “Programação e Fluxo Visual” com aproximadamente 350 steps/funções diferentes + plugins



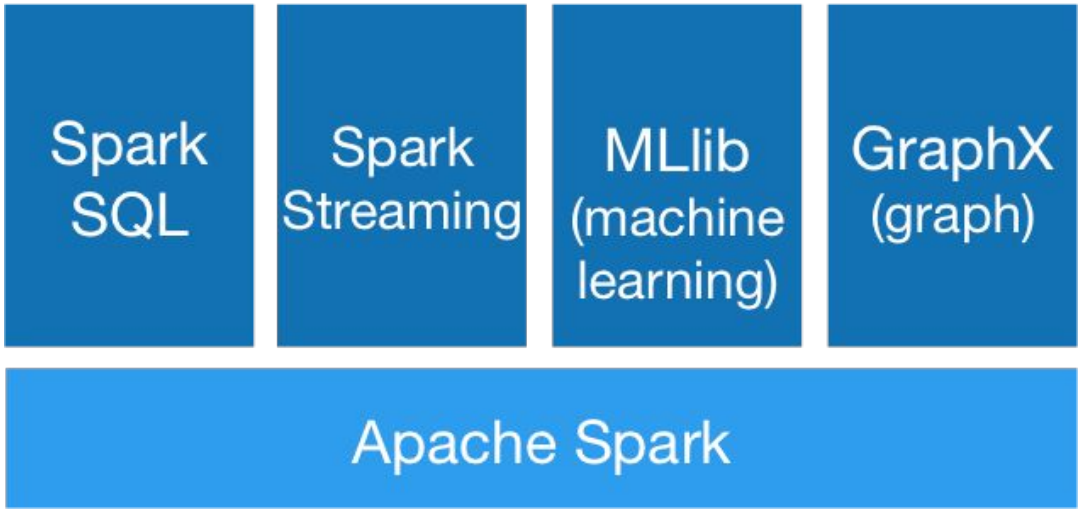


Modos de Execução do PDI



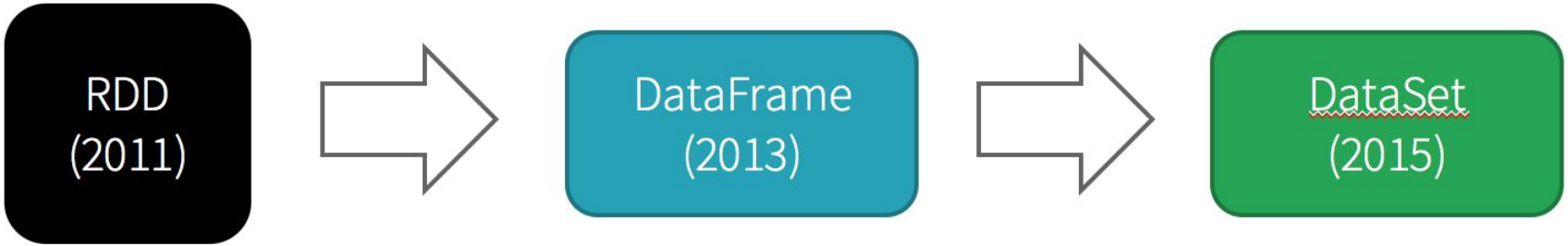
Quem é o Apache Spark?

- É um Mecanismo de análise unificada para processamento de dados em larga escala.



Evolução do Apache Spark

- Hoje podemos processar dados executando comando “SQL” no Spark, mas nem sempre foi assim!
- Cada vez mais conceitos estão sendo “encapsulados” e ficando mais faceis para os desenvolvedores



Como usamos Spark Tradicionalmente

- Carregamos datasets do HDFS, Hive, Cassandra, Hbase, Streaming, etc.
- Criamos RDDs e DataFrames.
- Processamos os Dados usando SparkSQL, Dataframes, Maps, Reduces, etc
- Os resultados gerar novos dados (Dataframes , RDDs etc)
- E segue o pipeline de processamento...

AEL - Adaptive Execution Layer

- O Pentaho usa a AEL) para executar transformações em **diferentes mecanismos**.
- Adapta etapas da transformação que você desenvolveu no PDI para operadores nativos no mecanismo selecionado para seu ambiente (Spark, Hadoop, Flink).
- O motor de Spark é mais adequado para a execução de grandes transformações de dados em um cluster Hadoop (**Hoje!**).
- Selecionando o mecanismo Spark para executar sua transformação, a AEL compara as etapas de sua transformação aos operadores nativos do Spark.

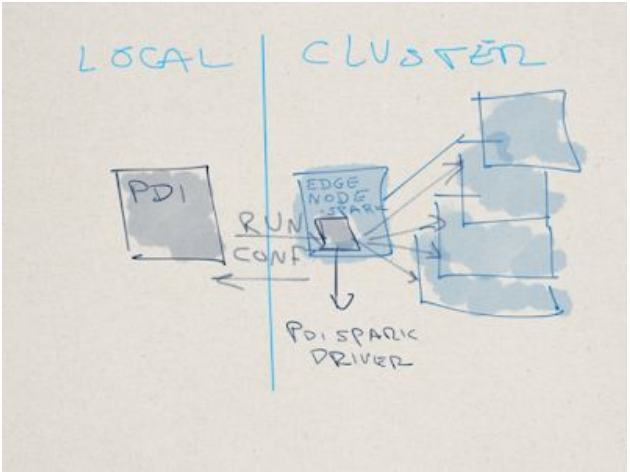
AEL - Adaptive Execution Layer

- Exemplo: se sua transformação contiver uma etapa de Entrada de Arquivo do Hadoop, a AEL usará um operador Spark equivalente.
- A AEL cria uma definição de transformação para o Spark, que move a execução diretamente para o cluster, aproveitando a capacidade do Spark de coordenar uma grande quantidade de dados em vários nós.

AEL – A ideia original

- Começamos por Driver PDI para Spark a partir de uma instância PDI, ponto de partida importante porque usando esta metodologia garantimos que qualquer plug-in que tenhamos desenvolvido / instalado funcionará quando executarmos a transformação (extensibilidade do Pentaho).

- O driver é instalado em um “edge node” do cluster responsável pela execução da transformação. Ao usar a spark, aproveitamos todas as suas características, podenddo executar **standalone** ou **yarn mode/ cluster**

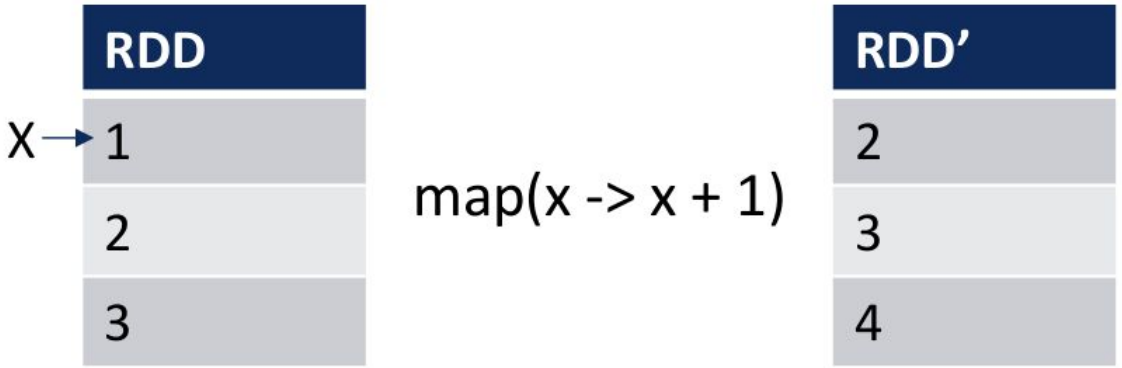


AEL

- Disponível no Pentaho deste Maio/2017 – já estamos na segunda onda de melhorias (agora Streaming, Kafka, etc) .

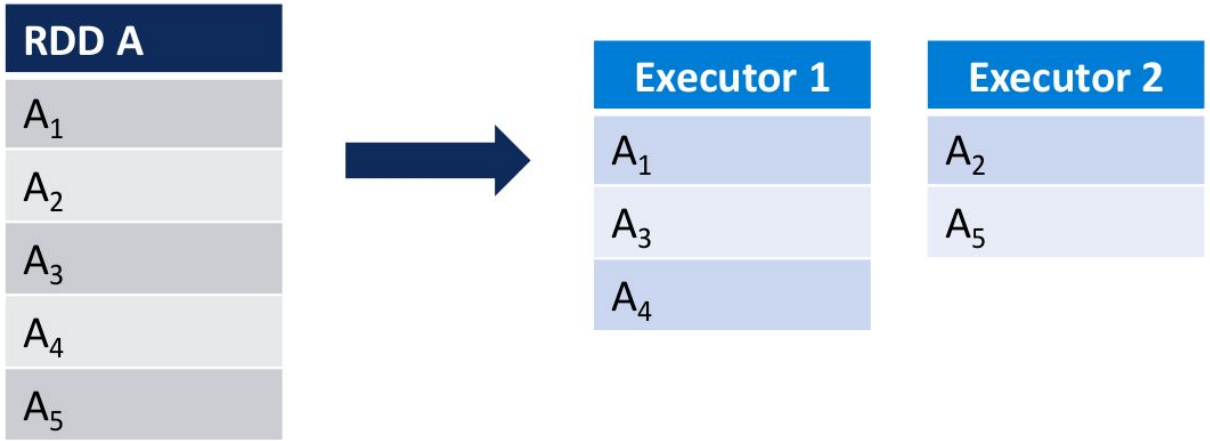
Spark - RDD

- O Spark trabalha com datasets chamado RDD
- Uma transformação (não é o mesmo que uma transformação PDI) descreve como produzir um novo RDD de um anterior:
- Uma ação executa um conjunto de transformações para produzir um resultado.



Spark - RDD

- O **executor** é o “trabalhador” no Spark
- Um RDD representa um dataset inteiro, que é dividido em **partitions**
- As divisões de arquivos do Hadoop geralmente definem o particionamento para RDD baseado em arquivo.
- Dados paralelizados pelo Spark Driver são divididos entre os executores

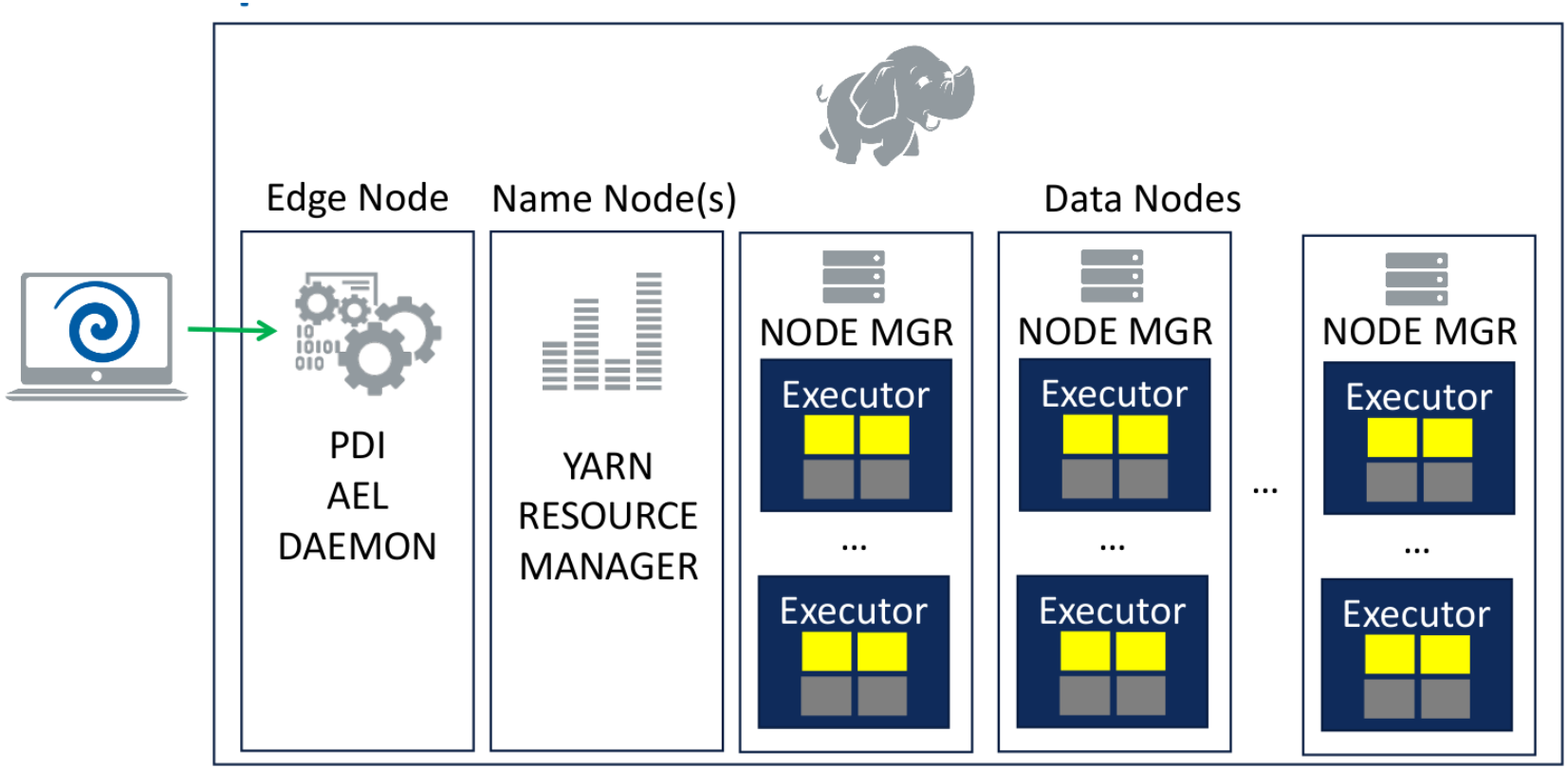


Spark - Tasks

- Representa a execução de transformações do Spark em uma partição de dados
- **Cores** especificam o número máximo de tarefas simultâneas a serem executadas
- Efetivamente, o número máximo de threads de processamento por executor

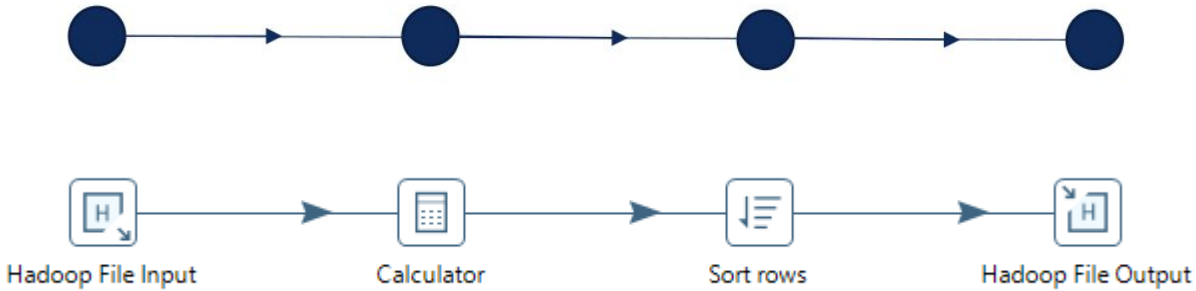
Tasks = 5, Cores = 2,
Total Max Threads = 4

AEL – Arquitetura – Visão Geral



Spark – Directed Acyclic Graph (DAG)

- Dirigido = Arestas podem ser percorridos em uma única direção
- Acíclico = Não há ciclos no grafos (você não pode visitar novamente um vértice)
- Grafos = Coleção de Vértices e Arestas



Transformações PDI como Spark DAGs

- Cada **step** executada como uma ou mais transformações do Spark
- A maioria dos **steps** são executados como transformações Spark
  - Utiliza o método processRow() do Kettle StepInterface
  - O mecanismo do PDI não executada no contexto do Spark, e sim como PMR (Pentaho MapReduce)
  - Muitas implementações processRow() existentes podem ser executadas em paralelo

PDI/Kettle	Spark
Transformation	DAG
Step	Transformation (1+)

RDD’ = RDD.mapPartitions(processRow())

Exemplo

- **Hadoop File Input** = Built-in Spark Input (Partitions data)
- **Calculator** = Generic Kettle Step (Partitioned)
- **Sort rows** = Built-in Spark Sort
- **Hadoop File Output** = Built-in Spark Output

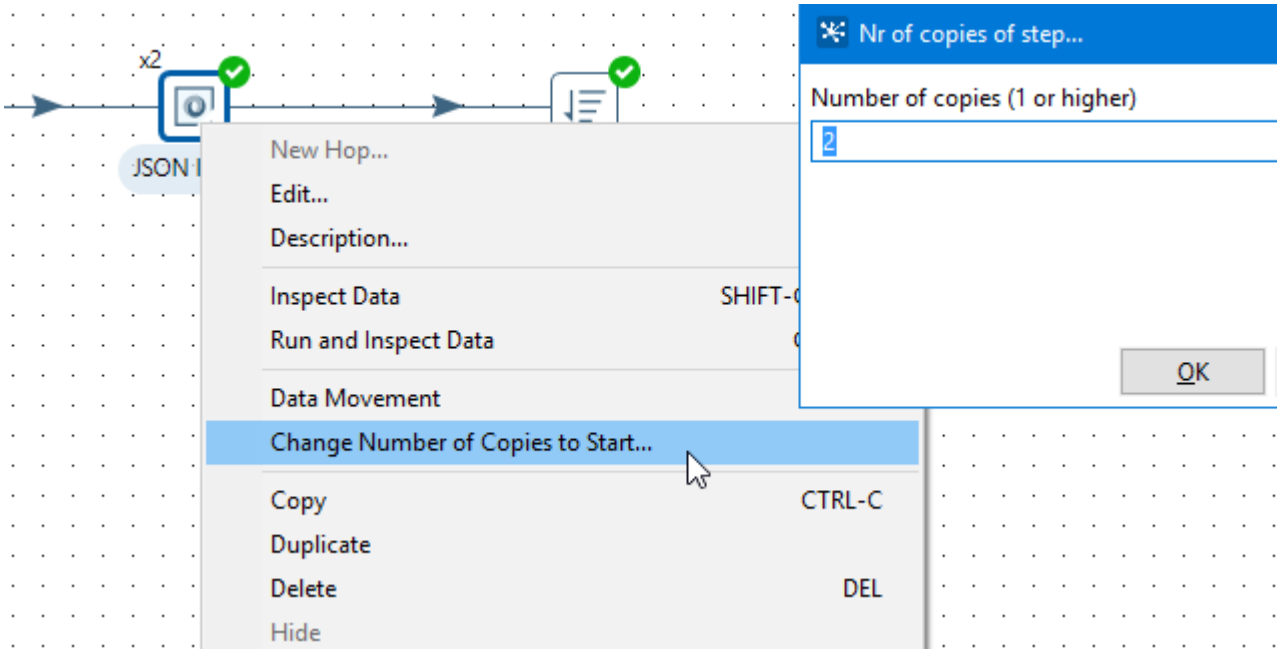

```
SparkContext.textFile(Input Files)
    .mapPartitions(Row Conversion)
    .mapPartitions(Generic(Calculator))
    .mapPartitions(Row Conversion)
    .sort(Sort rows metadata)
    .mapPartitions(Row Conversion)
    .saveAsTextFile(Output Files)
```

AEL – Kettle Engine

- Cada step é executado dentro de uma thread
  - As threads chamam um metodo chamado processRow()
- Ao executar, as linhas são roteadas pela transformação

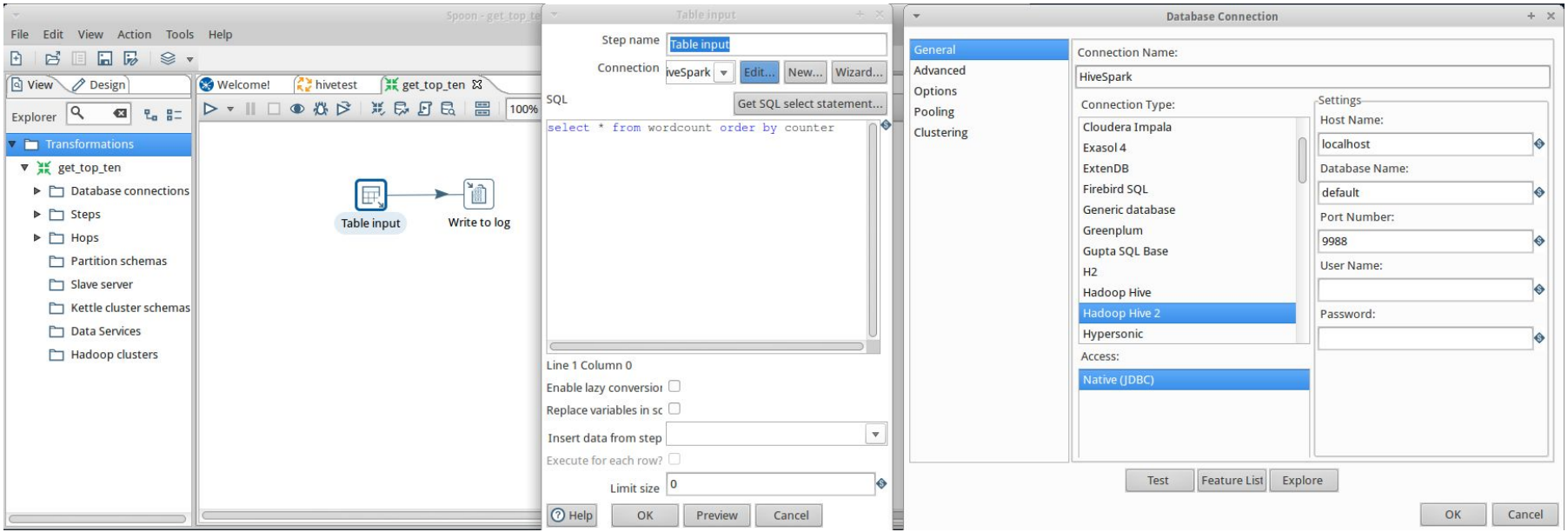
AEL – Kettle Engine - Paralelismo

- Podemos configurar o numero de copias para inicialização



Querys simples e implementação fácil

- Cloudera usa **Hive on Spark** com Hive2
- Hortonworks use **SparkSQL** via Simba

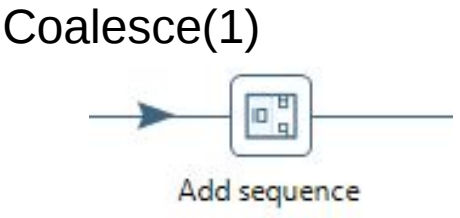


Steps que não tem paralelização

Dataset	Input	Output
Partition 1	A B	A, 1 B, 2
Partition 2	C D	C, 1 D, 2

Dataset Coalescing

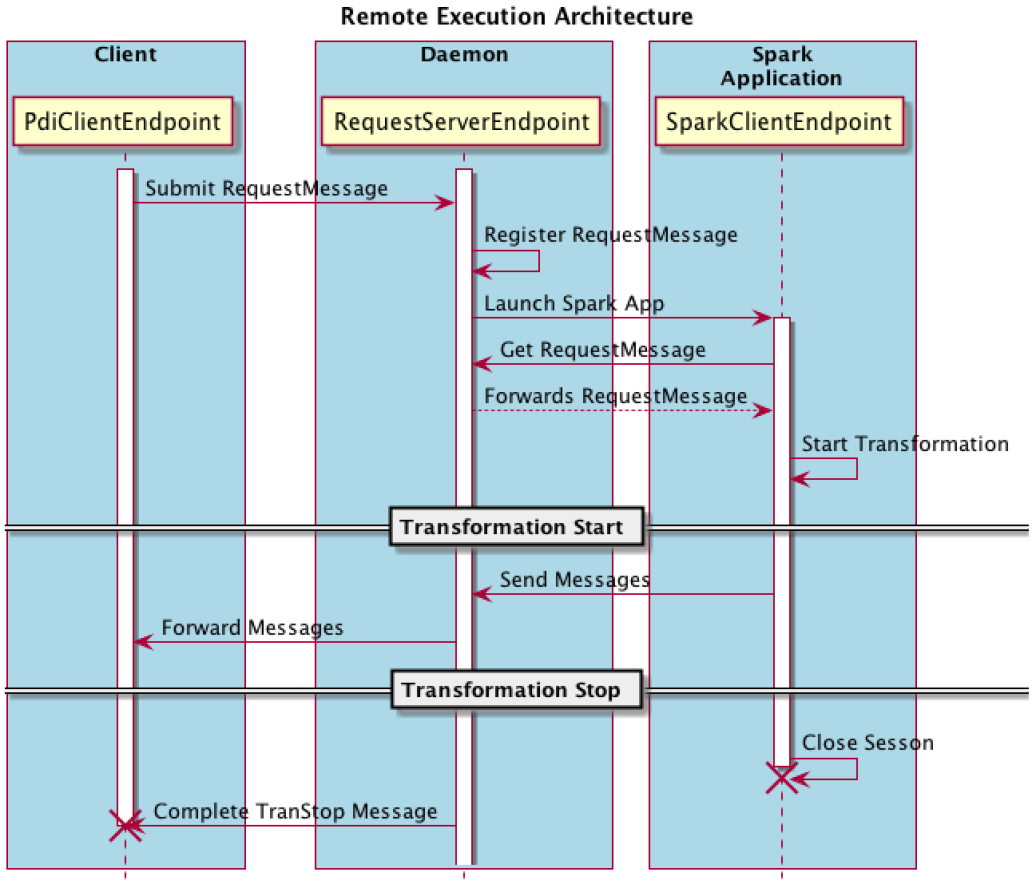
- A operação de **Coalescing** pode ser usada para mesclar as partições de um RDD
- Alguns ids de etapa são definidos na lista **forceCoalesceSteps**
- Configurável em org.pentaho.pdi.engine.spark.cfg



Dataset	Input	Output
Partition 1	A B C D	A, 1 B, 2 C, 3 D, 4

AEL – Arquitetura e Tecnologias

- Implementação em Spark
- O que é o Daemon?
  - Servidor web Lightweight Spring Boot
  - Orquestra as transformações em execução
- O que temos do Daemon?
- - Karaf / OSGi
  - Spark Modes
  - Rede DMZ
  - Menos configurações para o desenvolvedor do PDI



Websockets (JSR 356)

- **Uso de padrões**
- comunicação de 2 vias
- Interface e API orientada para anotações
- Mensagens leves
- Protegido pelo Spring Security (JAAS / Kerberos)
- Pode ser balanceado a carga
- Expansível para futuros recursos

Segurança

- SSL (Secure Sockets Layer)
  - Criptografar dados enviados pela rede

- Kerberos
  - Mecanismo de autenticação
  - Alavancar o KDC (Key Distribution Center) existente
  - Servidor Spoon / Pentaho para Daemon & Daemon -> Spark

- Proxy User
  - Usado quando conectado ao servidor Pentaho
  - Permite um usuário Pentaho executar **jobs** como esse usuário.

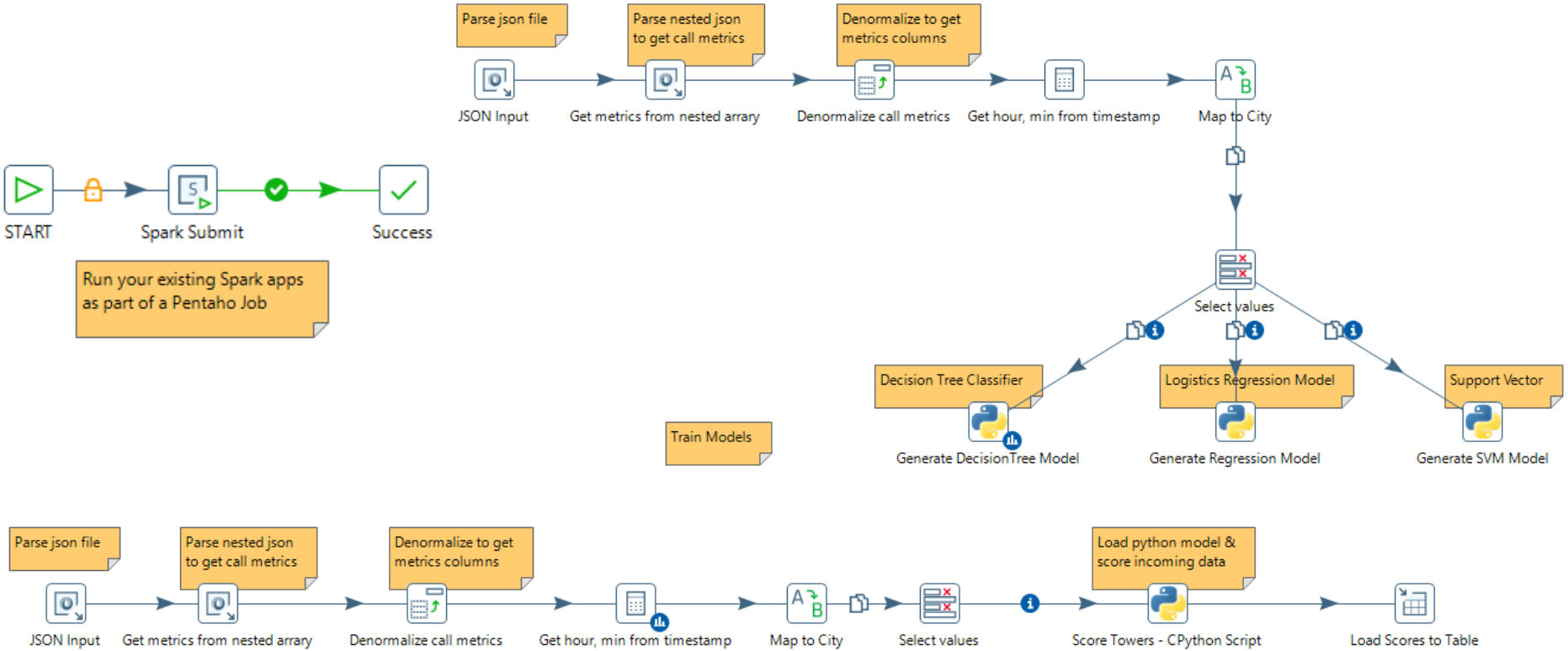
- Configuração simplificada

Compatibilidade e Suporte da AEL (Pentaho 8.0)

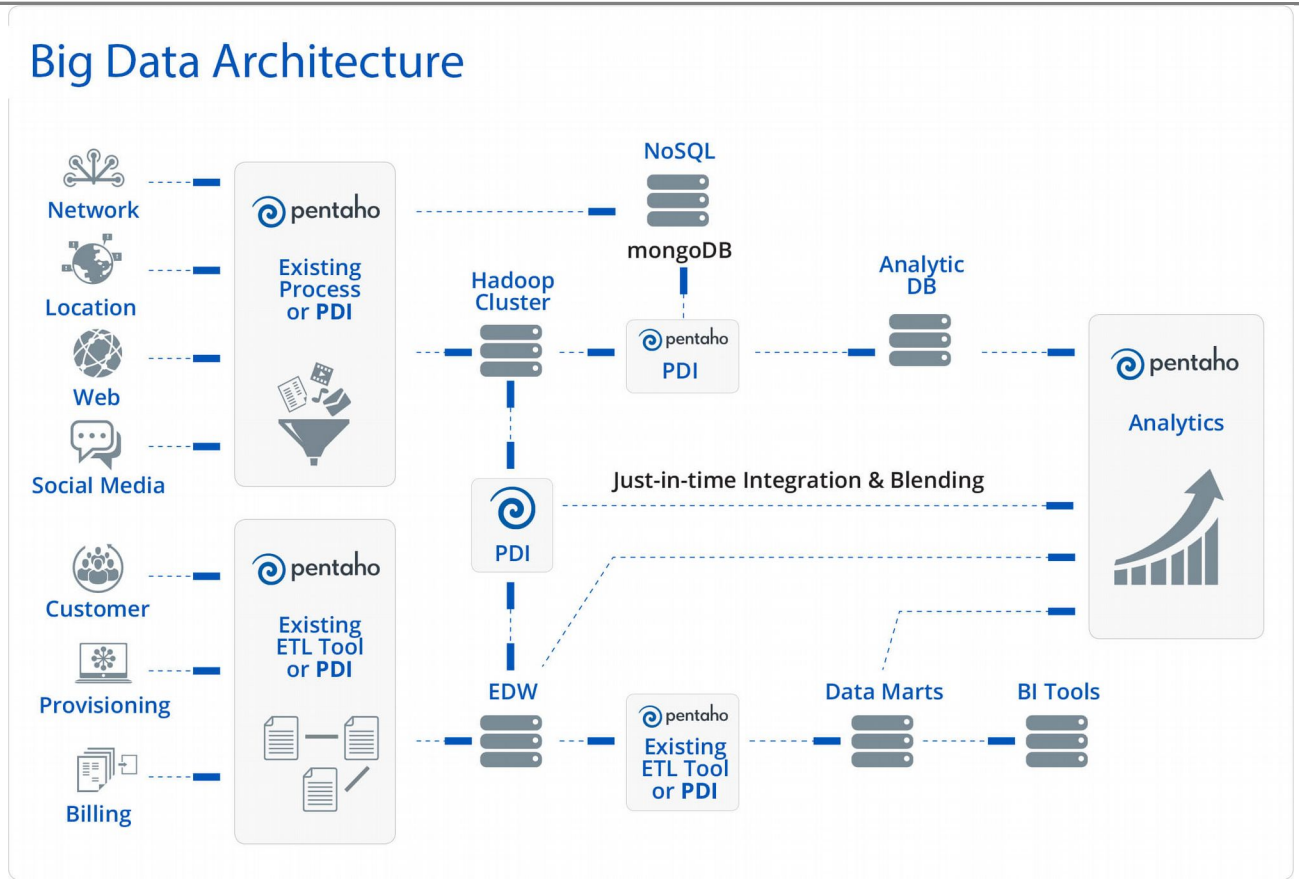
- Cloudera version 5.10 ou superior
- Hortonworks 2.5 ou superior
- MapR Spark 2.1
- Ou seu **Cluster 100% Apache!**
- Spark Client 2x.
- Microsoft Azure HD Insights shim
- Ranger support
- Kerberos Impersonation (Hortonworks) – Somente na EE
- Spark Streaming
- Kafka

Machine Learning e Data Science inclusive R, Python, WEKA com Spark

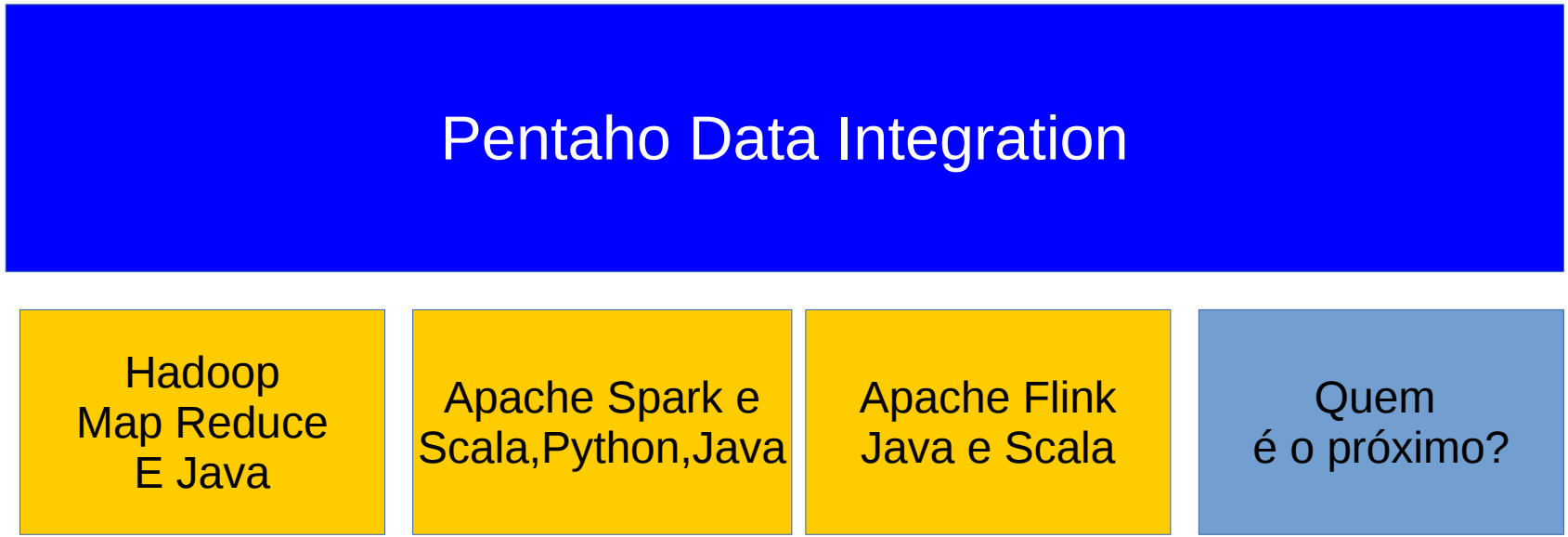
- Automação e orquestração do fluxo



Orquestração total com Pentaho



Resumo AEL

- 
Pentaho Data Integration
Hadoop
Map Reduce
E Java
Apache Spark e
Scala,Python,Java
Apache Flink
Java e Scala
Quem
é o próximo?

Setup AEL

- Step-by-Step
  https://help.pentaho.com/Documentation/8.0/Setup/Configuration/Adaptive_Execution_Layer
- Download Pentaho Data Integration
  https://sourceforge.net/projects/pentaho/files/Pentaho%208.0/client-tools/
- Download Apache Spark
  http://spark.apache.org/downloads.html
- Download Pentaho Shmis
  https://sourceforge.net/projects/pentaho/files/Pentaho%208.0/shims/

Referências

- **Pentaho World 2017**
  - Dev-by-Dev: PDI Engine and Adaptive Execution, Under the Hood
  - Design Patterns Leveraging Spark in PDI
  - Parallelism with the PDI AEL Spark Engine

- **Blog Pedro Alves -**
  http://pedroalves-bi.blogspot.com.br/2017/05/pentaho-7.1.html

- **Pentaho Day 2017**.
  Pentaho 7, Visão e Roadmap

MBA+

- Pedro está feliz de novo!
- E voltou a seguir Matt!
- É usuário Pentaho fiel!
- Pedro e um Data Scientist