# Pentaho

**Lasse Hoff Byrnak**
Solution Engineering - Big Data, Analytics & IoT
11/9-2019

# Agenda

Pentaho

Data sources

Demo

Advanced Analytics

Demo

# Our world

## BIG DATA & AI LANDSCAPE 2018

### INFRASTRUCTURE

HADOOP ON-PREMISE · HADOOP IN THE CLOUD · STREAMING / IN-MEMORY

NoSQL DATABASES · NewSQL DATABASES · GRAPH DBs · MPP DBs · CLOUD EDW

DATA TRANSFORMATION · DATA INTEGRATION · DATA GOVERNANCE · MGMT / MONITORING

STORAGE · CLUSTER SVCS · APP DEV · CROWD-SOURCING · HARDWARE · GPU DBs

### ANALYTICS

DATA ANALYST PLATFORMS · DATA SCIENCE PLATFORMS

BI PLATFORMS · VISUALIZATION · MACHINE LEARNING

COMPUTER VISION · HORIZONTAL AI · SPEECH & NLP

SEARCH · LOG ANALYTICS · SOCIAL ANALYTICS · WEB / MOBILE / COMMERCE ANALYTICS

### APPLICATIONS – ENTERPRISE

SALES · MARKETING - B2B · MARKETING - B2C · CUSTOMER SERVICE

HUMAN CAPITAL · LEGAL · FINANCE · ENTERPRISE PRODUCTIVITY · BACK OFFICE AUTOMATION · SECURITY

### APPLICATIONS – INDUSTRY

ADVERTISING · EDUCATION · GOVERNMENT · REAL ESTATE · FINANCE - INVESTING · FINANCE - LENDING · INSURANCE

HEALTHCARE · LIFE SCIENCES · TRANSPORTATION · AGRICULTURE · COMMERCE · INDUSTRIAL · OTHER

### CROSS-INFRASTRUCTURE/ANALYTICS

### OPEN SOURCE

FRAMEWORK · QUERY / DATA FLOW · DATA ACCESS · COORDINATION · STREAMING · STAT TOOLS · AI / MACHINE LEARNING / DEEP LEARNING · SEARCH · LOGGING & MONITORING · VISUALIZATION · COLLABORATION · SECURITY

### DATA SOURCES & APIs

HEALTH · IOT · FINANCIAL & ECONOMIC DATA · AIR / SPACE / SEA · PEOPLE / ENTITIES · LOCATION INTELLIGENCE · OTHER

### DATA RESOURCES

DATA SERVICES · INCUBATORS & SCHOOLS · RESEARCH

# Pentaho

# Self-Service Big Data Analytics

**Data Engineering**

Cleanse · Conform · Shape · Transform · Ingest

**Data Preparation**

Refine · Orchestrate · Virtualize · Blend · Prepare · Enrich

**Analytics**

Build · Score · Model · Visualize · Analyze

## Managing and Automating the Pipeline

| Administration | Security | Lifecycle Management | Data Provenance | Dynamic Data Pipeline | Monitoring | Automation |
|---|---|---|---|---|---|---|
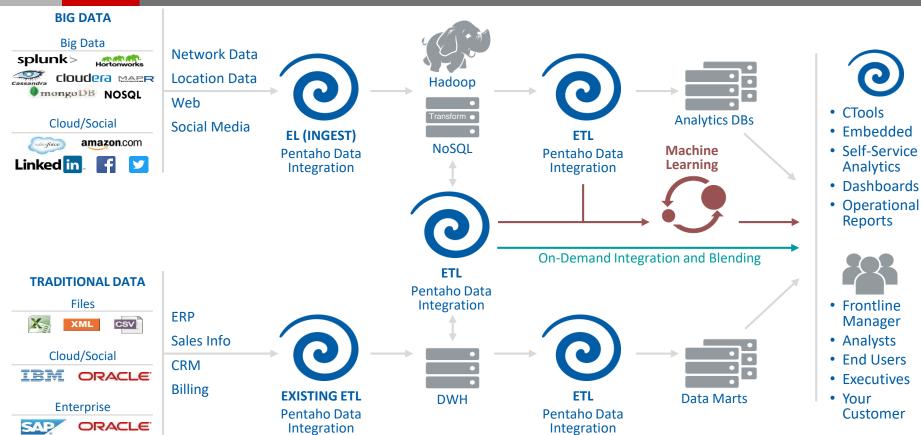
# Pentaho Brings Data Together
## Platform Supports Evolving Big Data Architectures
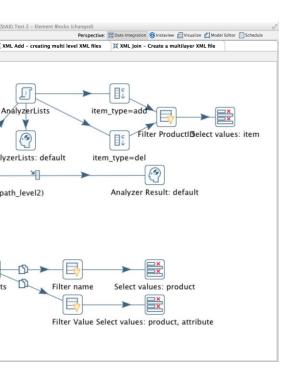
**BIG DATA**

**Big Data**
splunk>
Hortonworks
Cassandra   cloudera   MAPR
mongoDB   NOSQL

**Cloud/Social**
salesforce   amazon.com
Linked in   Facebook   Twitter

Network Data
Location Data
Web
Social Media

**EL (INGEST)**
Pentaho Data Integration

Hadoop

Transform

NoSQL

**ETL**
Pentaho Data Integration

Analytics DBs

**Machine Learning**

**ETL**
Pentaho Data Integration

On-Demand Integration and Blending

- CTools
- Embedded
- Self-Service Analytics
- Dashboards
- Operational Reports

**TRADITIONAL DATA**

**Files**
XML   CSV

**Cloud/Social**
IBM   ORACLE

**Enterprise**
SAP   ORACLE

ERP
Sales Info
CRM
Billing

**EXISTING ETL**
Pentaho Data Integration

DWH

**ETL**
Pentaho Data Integration

Data Marts

- Frontline Manager
- Analysts
- End Users
- Executives
- Your Customer

6

## What is PDI?

- Graphical user interface to create drag-and-drop data transformation pipelines

- Data integration engine that is powerful, portable, and scalable

- Software that can be easily extended and embedded into other applications

# Bringing Analytics into Data Prep

Spot check and visualize data in flight – without switching applications

# Data sources

# Support a lot of data sources
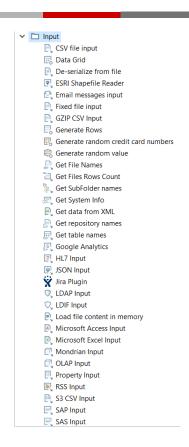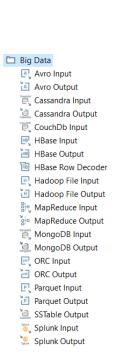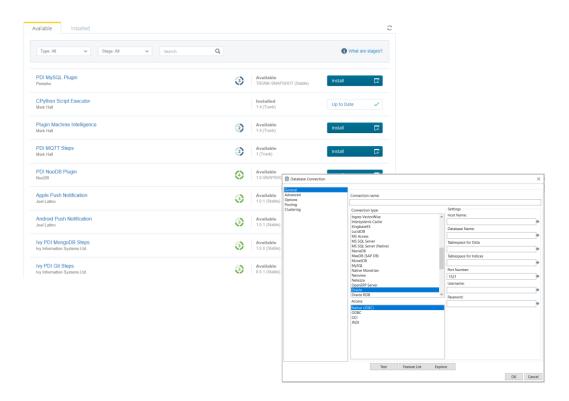
# Data Sources

Apache Kafka Consumer

CSV file input

MongoDB Input

Table input

Text file input

IC AMQP Plugin

MQTT Consumer

JSON Input

Splunk Input

REST Client

Hadoop File Input

HTTP Client

**Demo**

# Simple transformation

Read Sales Data → Filter rows → Value Mapper → Number range → Write to Database
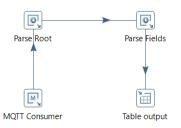
Read Postal Codes → Lookup Missing Zips → Prepare Field Layout

# Advanced Analytics

# Advanced Analytics

# Advanced Analytics

**Demo**

# Monitor Model

1. Data and feature engineering

Join and blend data from different sources.

input - customer billing

input - customer transaction

input - transaction_fraud_report

input - customer 2 → lookup - customer → lookup - transaction_details → lookup - transaction_fraud_report → billing_shipping_zip_equal → Select values

In this example, we just use all the fraud data for monitoring. In a real-world scenario, some amount of the most recently labeled historical data would be used to monitor performance of the static models.

PMI Scoring    PMI Scoring 2    PMI Scoring 3    PMI Scoring 4    PMI Scoring 5

| Current champion | Weka Decision T | R Logistic Reg | sklearn GBT | MLlib NB |
|---|---|---|---|---|

Replace in string

Monitoring Output

Monitoring output CSV file could be a data service
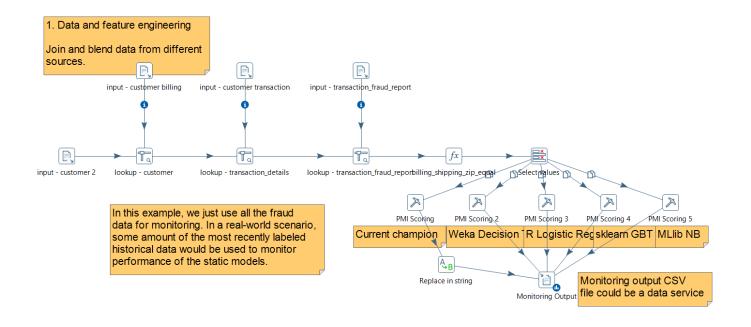
# Resources

**Community Edition**
https://sourceforge.net/projects/pentaho/

**Enterprise Edition (30 days trial)**
https://www.hitachivantara.com/en-us/products/big-data-integration-analytics/pentaho-trial-download.html

**Pentaho Community forum**
https://community.hitachivantara.com/s/pentaho

**Pentaho @ github**
https://github.com/pentaho

Thank You