Exploring Pentaho’s Role in IoT Data Possibilities

Anjali Rajith

LinuxCon Japan 2016
July 15th, 2016
Tokyo

Center of Technology Innovations – System Engineering,
Hitachi Ltd., Research & Development Group,
Yokohama-shi, Kanagawa-ken, Japan
E-mail: anjali.rajith.he@hitachi.com
This presentation aims to open the door to a highly efficient open source data integration and analysis platform, Pentaho, with seamless big data / IoT data analysis capabilities.

It aims to discuss how Pentaho stands out from the crowd by continuously evolving with native integration of trending open source technologies.
Contents

• Introduction to IoT & IoT Data
• Overview of business analytics
• Introduction to Pentaho & it’s components
• IoT data management using Pentaho
  ➢ Big data environment
  ➢ Data mining capabilities
  ➢ Interactive reporting
• Case studies
• Conclusion
IoT – Internet of Things

- Too vast to be defined, as it’s possibilities are endless!
- IoT creates a connected and closed world.
  - devices and equipment connected to each other with sensors, software & network to collect and exchange data.
What is IoT Data?

• Automatically-collected machinery data, weather data etc.
  ➢ They contain sensor information such as temperature, signal quality etc.
• Complex, vast & rapidly moving.
• Utilizing IoT data helps to improve operational quality and reduce cost.
• Analysis of sensor data helps to
  ➢ Spot anomalies in machines.
  ➢ Remote machine monitoring.
Automation in IoT

- Automated IoT data collection, integration, classification and analysis helps to make right decisions at right time.
- Real-time IoT data analysis helps to uncover correlations between current data and historic data.

Data collection
Data storage
Data cleansing
Data transformation
Data analysis
Data prediction
Business Analytics

• Integration and statistical analysis of organization’s data for timely decision making.
  ➢ Accessible anytime from anywhere by anyone.
  ➢ The business insights after analysis help in automation and optimization of processes.
  ➢ Predicts failures.
  ➢ Reduces cost and improves operational efficiency.
IoT Data Analysis applications in various sectors

- Automated farm monitoring
- Ship information system
- Machinery Anomaly detection
- Automated traffic monitoring
Future IoT World!

- All devices are inter-operable!!
IoT Challenges

• 54% of organizations’ IoT data analysis are inefficient.
  ➢ Analytical tools are inadequate and scattered.
  ➢ Inability to leverage data in semi-structured & unstructured format.
  ➢ Inability to convert the accumulating data into insights.
  ➢ Inability to handle the rapid inflow of data to make quick decisions.
  ➢ High cost.
Birth of an Open Source Tool!

- Pentaho - founded in 2004 at Orlando, USA.
- A Hitachi-group company from 2015.
- Pentaho BI project is a pioneering initiative by open source community which provides
  - Pentaho Business Analytics, a suite of comprehensive Data Integration and Business Intelligence platform.
    - Gather, integrate and analyze big & IoT data.
  - provides organizations with business intelligence capabilities to improve their business efficiency and performance.
Pentaho

• Covers all operations from data integration to data mining.

• Easy-to-use with seamless integration of variety of data sources.

• Processes data in big data environment.
  ➢ Native integration with Hadoop, Spark, NoSQL and other analytical databases.

• Powerful data mining capabilities using R models & Weka tool.

• Web-based user interface & interactive reporting.
Overall Insight into Pentaho

Structured Data

Un-structured Data

Pentaho

Data Integration
ETL (Extract Transform & Load)

Data Discovery
Reports
Dashboards
analysis

Data Mining
Predictive Analysis & Machine Learning

Business Insights

© Hitachi, Ltd. 2016. All rights reserved.
Pentaho Main Components

• Servers: Data Integration (DI) & Business Intelligence (BI) server

• Design tools
  ➢ Spoon – GUI to perform ETL.
  ➢ Design studio – Collection of editors to create action sequences.
  ➢ Schema Workbench – OLAP modelling.
  ➢ Report Designer – Create reports.
  ➢ Enterprise console/Admin console – add credentials, monitor performance, remote monitoring, verify connections etc.

• Various BI plugins and DI plugins
  ➢ Big Data & Marketplace
Pentaho Data Integration (PDI)

• Java-based ETL engine - helps to load data into external databases. (Extract Transform Load)
  ➢ Migrates data massively
  ➢ Integrates and transforms data
  ➢ Data cleansing
  ➢ Connects to variety of data sources- databases, data warehouses, web APIs & data files
  ➢ Loads data to target databases such as data warehouses, data lakes etc.
Spoon - GUI to create transformations or jobs

Pan
Perform read/write/manipulate Data to & from data sources

Kitchen
Launch jobs

Repository

PDI server
Runs jobs / transformation
Pentaho Server High Availability (HA) Solution

- Pentaho provides HA – clustered solutions for large-scale deployment.
  - Handles increasing data and concurrent users.

- Multiple Pentaho nodes form a cluster.
  - Each node contains a Tomcat Web App server and a BA server.
  - A load balancer (SSL) helps to spread computing resources evenly among nodes.
PDI clustering

Dynamic clustering at run-time!

Data from various sources

Distributes ETL workload

Load balancer

Master

Slave1 + PDI

Slave2 + PDI

Slave3 + PDI

Slave4 + PDI

Target
Pentaho Analysis

• Pentaho Analysis Engine, Mondrian OLAP creates analysis schema.
  ➢ presents results in multi-dimension format for OLAP analysis using MDX query.

• In-memory caching.
  ➢ Faster ad-hoc analysis of data.

• In-memory aggregation.
  ➢ Rolling up of granular data in in-memory.
Slice & Dice
Drill through the cube

SELECT {[Measures].[sales]} ON COLUMNS,
{[Product].members} ON ROWS
FROM [Sales]
WHERE [Time].[2013].[Q2]
Pentaho Business Analytics

• Pentaho Business Analytics server is a web application for creating and sharing interactive reports and dashboards.
  ➢ dynamic reports in various formats - pdf, html, excel, csv etc.
  ➢ Can be directly connected to PDI to receive data real-time.
Pentaho Web Interface
Pentaho Business Analytics

Users
- Data warehouse developers
- Organizations / business users
- Data analysts / data scientists

Pentaho Business Intelligence
- Interactive Web interface
- Attractive dashboards
- Reports
- Predictive analysis

Pentaho suite
- Generate reports directly
- Data Integration and Transformation

Data sources

© Hitachi, Ltd. 2016. All rights reserved.
Pentaho Design Studio

• Eclipse application framework that provides a graphical environment to build action sequences.

  ➢ Action sequences / X-actions define activities ranging from database queries, PDI jobs, report generations, email actions and the order they should occur.

```
Action Sequence1
  └── Write database query
        ├── Add parameters
        └── Create Report
            └── Send email
```

Action Sequence2

7/7/2016
Pentaho’s Support for Big Data

- Pentaho empowers users by blending with the state-of-the-art big data technologies for quick and efficient data analytics.

Analytical databases
(Teradata, Netezza, Greenplum etc.)

Spark

Hadoop

NoSQL
(Cassandra, HBase, MongoDB etc.)
Pentaho + Hadoop

• Pentaho is integrated to Hadoop –
  ➢ Traditional ETL – HDFS files, Hbase, Hive & Impala queries Read and Write.
  ➢ Data Orchestration – HDFS copy files, Map Reduce jobs execution, Pig Script execution etc.
  ➢ Pentaho Map Reduce – Graphical designer to visually build MapReduce jobs and run in cluster.
Pentaho + Hadoop

Pentaho DI

Edge node

Hadoop cluster

Pentaho Map Reduce

MapReduce input → Regex Evaluation → Filter rows → Value Mapper → User Defined Java Expression → MapReduce Output
• Spark was introduced by Apache as an open source cluster computing framework for high speed data processing.
  ➢ 100X faster than Hadoop in-memory
  ➢ 10X faster on disk

• Speeds up Hadoop computational computing software process.

• Covers batch applications, iterative algorithms, interactive queries and streaming.

• High speed big data processing platform.
Pentaho + Spark

• Pentaho is natively integrated with Spark!
• Develop solutions using Spark in Pentaho.
• Pentaho’s open source approach enables it to evolve along with the new technologies.

Mlib is Spark’s scalable machine learning library

Advanced data analytics

Pentaho

+

Apache Spark + Mlib
Pentaho + Spark
IoT Data Analysis with Pentaho

- Extracts, transforms & loads streaming data from machine sensors installed on customer equipment.
- Stores data in big data platforms like Hadoop.
- Weka, R & Python plugins in PDI helps to model and visualize different usage patterns and habits.
  - RScriptExecutor
  - Weka Scoring
  - PDI Python Integration
  - Mlib Spark library
IoT Data Collection

Sensor Controller management

Sensor management function

Gateway | Sensor
---|---

Information gather request

Data analysis

(Pentaho Reports/Dashboards)

Pentaho Data Integration

IoT data protocol (MQTT)

Data Prediction

Railway

Gateway

NX Dlink (ISO 15745-4)

Factory

Gateway

OPC UA

Power plant

Gateway

IEC61850

Sensor Controller management
Pentaho + R

- Execution of R Scripts in PDI via plugins.
- It’s very simple!
- Loads R scripts run-time.
  - Install R, set environmental variables & configure spoon with rJava.

PDI
Data integration from various sources

+ R
Handles statistics and numbers in the data

Valuable Data insights & prediction
Weka (Waikato Environment for Knowledge Analysis)

• Suite of Machine learning software written in Java.
• Contains a collection of visualization tools and algorithms + graphical interface.
• Supports data mining tasks such as data preprocessing, clustering, classification, regression and feature selection.
Weka Interface
Weka Scoring Plugin in PDI

• IoT/big data can get really complex because of the diversity of data sources and analytics involved.
  ➢ Pentaho helps to simplify this task with an agile platform.

• Weka scoring allows to score data as a part of PDI transformation by applying classification, clustering, and regression models constructed in Weka.
Weka Scoring Plugin in PDI

1. Read credit risk data.
2. Score using the rpart decision tree.
3. Convert probability of good risk to a percentage value.
4. Select the fields we want.
Pentaho + Hadoop + Weka

The ever-increasing volume of data growing demand for Hadoop machine learning algorithms to derive value.

Data \rightarrow \text{map} \rightarrow \text{map} \rightarrow \text{map} \rightarrow \text{reduce} \rightarrow \text{HDFS} \rightarrow \text{Pentaho}

Learn models
Remote Machine Monitoring

Pentaho suite

Data

Fast real-time reporting

BI

Factory

Gateway

OPC UA

7/7/2016

© Hitachi, Ltd. 2016. All rights reserved.
Data Prediction

Pentaho suite

Factory
Gateway
OPC UA

BI
predicted values

12:00 12:05 12:10 12:15 12:20 12:25 12:30

0 10 20 30 40 50

7/7/2016

39
Pentaho Marketplace

[Image of Pentaho Marketplace interface]

PDI MySQL Plugin
Pentaho

PDI NuoDB Plugin
NuoDB

Apple Push Notification
Joel Latino

Android Push Notification
Joel Latino

Ivy PDI MongoDB Steps
My Information Systems Ltd.

Ivy PDI Git Steps
My Information Systems Ltd.

Available: TRUNK-SNAPSHOT (Stable)
Available: 1.0-SNAPSHOT (Stable)
Available: 1.0.1 (Stable)
Available: 1.0.1 (Stable)
Available: 1.0.0 (Stable)
Available: 0.0.1 (Stable)

Install
Install
Install
Install
Install
Install
Case Studies

• **Halliburton Landmark – Oil & Gas**
  - Advanced analytics using Pentaho to monitor machine-sensor data.
    - So, improves pump safety and prevent spills by predicting failure rates.

• **Caterpillar Marine – maritime data analysis & remote monitoring technology.**
  - PDI for ETL & Pentaho Business Intelligence for reporting.
  - Weka predictive plugin to model patterns and predict machinery failure.
Case Studies

- Intelligent Mechatronics System – Automotive industry.
  - Use Pentaho to derive value out of big data in connected-car programs.
  - Usage-based insurance programs.

- Paytronix – provider of loyalty, gift and email solutions. Integrates big data to derive values for restaurants.
  - Use Pentaho to integrate data such as customer preferences, gift programs etc., perform ETL, data migration to Hadoop and other PDI functionalities.
Conclusion

- Pentaho helps organizations to anticipate business opportunities, take appropriate decisions in timely manner and reduce equipment failure and cost.
- The support of big data platforms helps to scale up along with data.
- Strong machine learning tools help to undertake necessary actions.
- The support for variety of data sources allays data compatibility issues and related problems.
Try Out

- Pentaho + Raspberry PI
- Pentaho + Docker

Pentaho Github:
https://github.com/pentaho

References:
www.pentaho.com
www.cs.waikato.ac.nz/ml/weka/
spark.apache.org/
Disclaimer

Pentaho, the Pentaho BI project, the Pentaho BI platform, and/or other Pentaho products and logos are within registered trademarks of Pentaho in the United States and other countries. We are not affiliated with, endorsed or sponsored by Pentaho / Pentaho community.

The names and logos of other companies or products may be registered trademarks of their respective owners.