Agenda

Pentaho’s Latest and Upcoming Features for Processing Big Data – Batch or Real-time

• Process big data visually in future-proof way
  – Demo

• Combine stream data processing with batch
  – Demo
Big Data Processing is HARD

1) Gartner Analyst, Nick Heudecker; infoworld.com, Sept 2015

"Through 2018, 70% of Hadoop deployments will not meet cost savings and revenue generation objectives due to skills and integration challenges." – GARTNER

1) Gartner Analyst, Nick Heudecker; infoworld.com, Sept 2015
Big Data Integration and Analytics Workflow with Pentaho

Big Data Challenges
- Processing Semi/un/structured data
- Blending big data with traditional data
- Maintaining security, governance of data
- Processing streaming data in real time and historically
- Enabling and operationalizing data science
Process Big Data Visually in a Future Proof Way
Visual Big Data Processing with Pentaho

• **What**: Visually ingest and process Big Data at enterprise scale

• **What Special**: Visually develop once and execute on any engine with Adaptive Execution Layer (AEL)

• **Why**
  – Difficult to find qualified developers
  – Difficult to keep up with new technologies

• **Available since Pentaho 7.1**
Adaptive Execution of Big Data

**Challenge:**
With rapidly changing big data technology, coding on various engines can be time-consuming or impossible with existing resources.

**Solution:**
Future-proof data integration and analytics development in a drag-and-drop visual development environment, eliminating the need for specialized coding and API knowledge. Seamlessly switch between execution engines to fit data volume and transformation complexity.
Adaptive Execution for Spark

Process Big Data Faster on Spark Without Any Coding

Challenge:
Finding the talent and time to work with Spark and newer big data technologies

Solution:
More easily develop big data applications in PDI using adaptive execution to ingest, process and blend data from a range of big data sources and scale on Spark clusters
Upcoming Enhanced Adaptive Execution Layer

- **Simplified Setup**
  - Fewer steps to setup
  - Easy to configure fail-over, load-balancing

- **Development productivity**
  - Robust transformation error and status reporting
  - Customization of Spark jobs

- **Robust Enterprise Security**
  - Client to AEL connection can be secured
  - End-2-end Kerberos impersonation from client tool to cluster
Upcoming Big Data File Format Handling

Big Data platforms introduced various data formats to improve performance, compression, and interoperability.

What:
• Visual handling of data files with Big Data formats Parquet and Avro
  – Reading and writing files with specific steps
  – Natively execute in Spark via AEL

Why:
• Ease of development of Big Data processing
• Performance improvement due to avoidance of intermediate formats
Demonstration
Retail Web Log Data Processing with Pentaho

- Run within Spoon via Pentaho during development and then use Spark cluster for production
- Lookups, sort, and Parquet file in/out and other steps as to test parallel and serial processing within Spark Cluster
Combine Stream Processing with Batch Processing
What is Stream Data Processing? And Why?

• Batch data processing is useful, but sometimes businesses need to obtain crucial insights faster and act on them

• Many use cases must consider data 2+ times: on the wire, and then subsequently as historical data

• Get crucial time-sensitive insights
  – React to customer interactions on a website or mobile app
  – Predict risk of equipment breakdown before it happens

Former POV “secure data in DW, then OLAP ASAP afterward”
gives way to

Current POV “analyze on the wire, write behind”
NEW Stream Data Processing with Pentaho

- Visually ingest and produce data from/to Kafka using NEW steps
- Process micro-batch chunks of data using either a time-based or a message size-based window
- Switch processing engines between Spark (Streaming) or Native Kettle
- Harden stream processing libraries and steps to process data from traditional message queues

**Benefits:**
- Lower the bar to build streaming applications
- Enable combining batch and stream data processing
How to Process Stream Data in Pentaho

• Steps for Kafka ingestion and publish
  – Kafka Consumer
  – Kafka Producer

• Steps for stream processing
  – Get records from stream

• Ingest and process continuous stream of data in near real-time in parent transformation

• Process micro-batch of stream data in separate child transformation
Combined Data Processing Using Spark & Pentaho

**DATA SOURCES**
- IoT Data
- Web Clickstream and Other Logs
- Traditional DB/DW and NoSQL Datastores
- Traditional Message Bus

**Micro Services**
- RT Data Processors
- Batch Data Processors

**HADOOP/SPARK CLUSTER**
- Spark
- Hadoop MR
- HDFS

**Pentaho DI**
- Collects data from sources including Kafka Clusters
- Can process streaming data using Spark and Spark Streaming or Kettle engine in a completely visual way
- Can retrieve processed or blended data from Hadoop/Spark and publish to Kafka clusters or external databases

**Ingest**

**Process**

**Publish**

**Reporting**

**Kafka Cluster**
Demonstration
Retail Store Event Processing

• Can be run within Spoon via Pentaho or within AEL-Spark engine

• Utilizes Kafka in/out, Parquet out and other steps as to demonstrate stream data ingestion, window processing and much more…
Availability and Roadmap
Availability

• Adaptive Execution Layer (AEL) and Spark-AEL available in Pentaho 7.1
  – Secure Spark integration, high-availability and security of AEL is EE only
  – Supported Hadoop distros in Pentaho 7.1 - Cloudera CDH and Pentaho 8.0 – Cloudera CDH and Hortonworks HDP

• Kafka steps and stream data processing available in Pentaho 8.0
  – Kafka from Cloudera and Hortonworks to be supported
Roadmap

- Extending AEL to support other Spark distros and other data processing engines
- Advanced stream processing with other real-time messaging protocols and windowing mechanism
- Enabling Big Data driven machine learning on batch or stream data
- Integrated with broader Hitachi Vantara portfolio
SUMMARY: Visual Future-Proof Big Data Processing with Pentaho

Leverage the power of Adaptive Execution to future-proof data processing pipelines
- Configure logic without coding
- Switch processing engines without rework
- Handle Big Data formats more efficiently

Visually build stream data processing pipelines for different streaming engines
- Configure Stream data processing logic
- Execute logic in multiple stream processing engines without rework
- Connect to streaming data sources

NEW in Pentaho
- Adaptive Execution Layer
- Visual Spark via AEL
- Native Big data Format Handling

NEW in Pentaho
- Native Streaming in PDI
- Spark Streaming via AEL
- Kafka Connectivity
Next Steps

Want to learn more?

• Meet-the-Experts:
  – Anthony DeShazor
  – Luke Nazarro
  – Carlo Russo

• Recommended Breakout Sessions:
  – Jonathan Jarvis: Understanding Parallelism with PDI and Adaptive Execution with Spark
  – Mark Burnette: Understanding the Big Data Technology Ecosystem