HITACHI
Inspire the Next

Pentaho Data Integration

Pentaho Analytics Platform

James O'Reilly

Date

Module Objectives

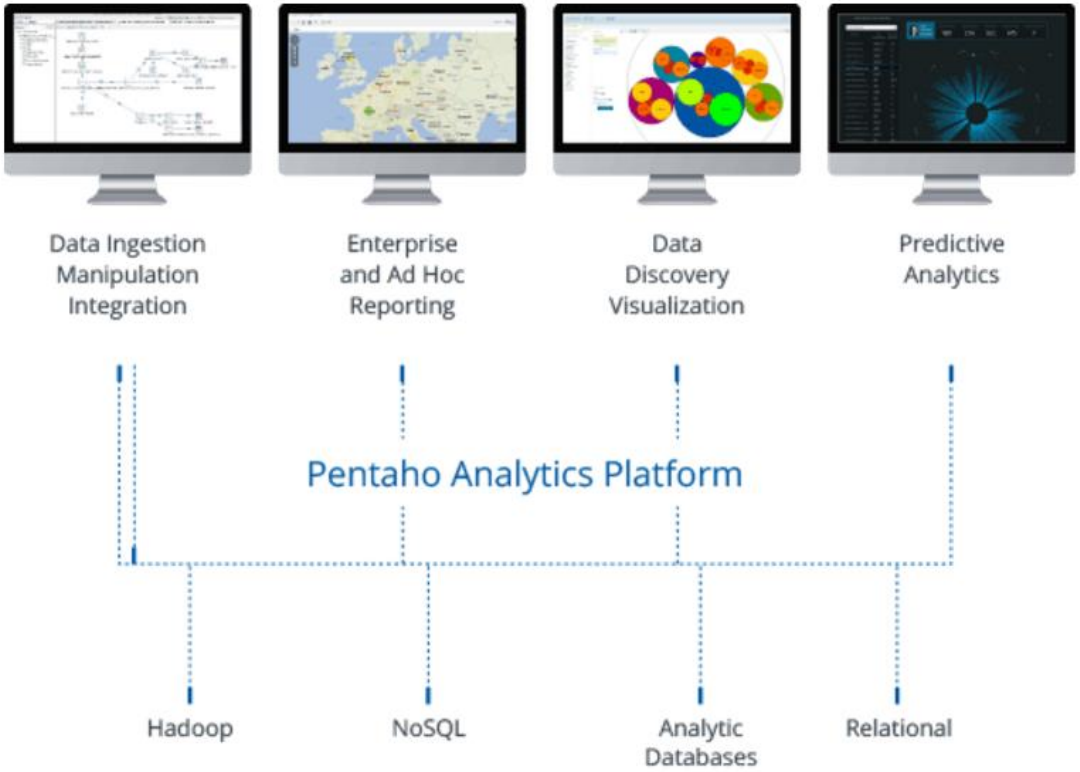
HITACHI
Inspire the Next

When you complete this module, you should be able to:

- Outline the Pentaho Platform
- Configure & Describe key Pentaho Data Integration components
- Understand the concept of Parallelism
- Define Transformations

Pentaho Analytics Platform

HITACHI
Inspire the Next



Web-Based Tools and Plugins

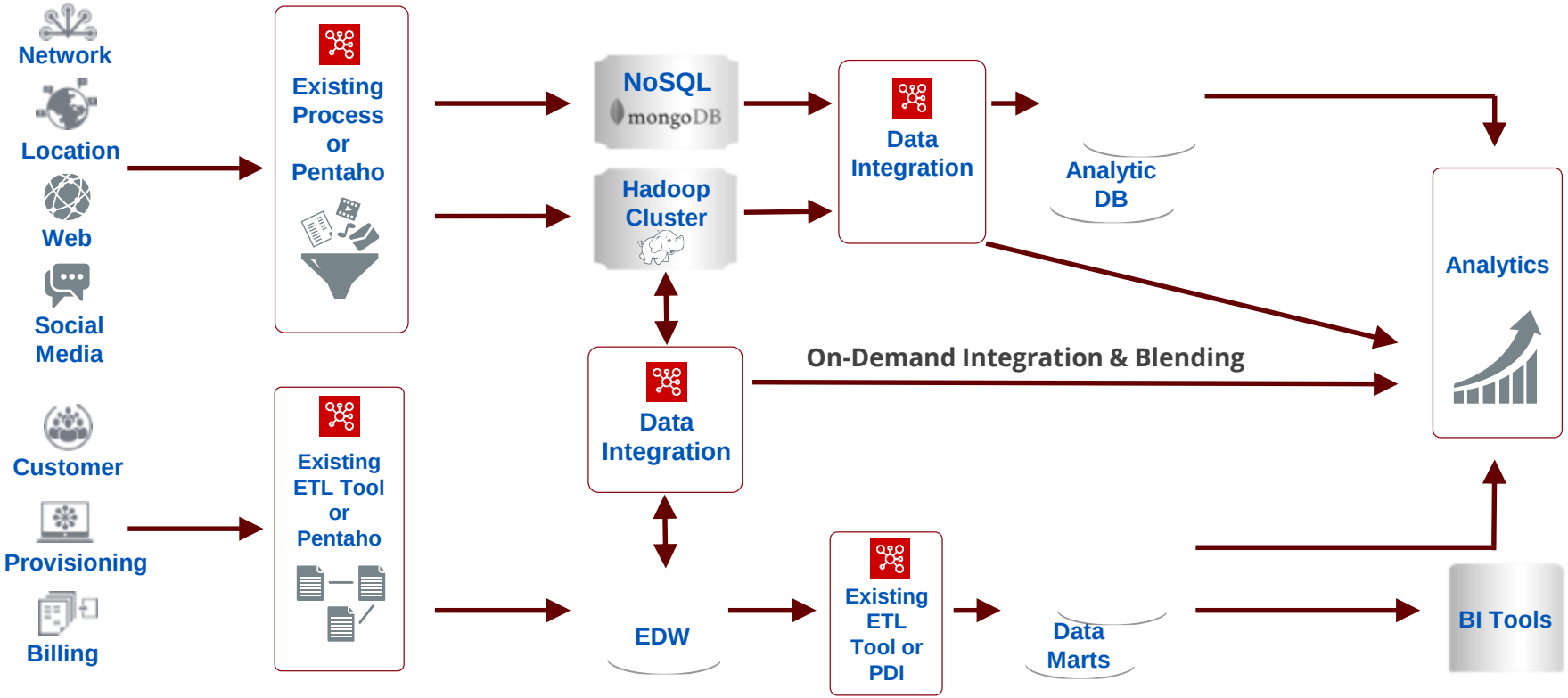
- Analyzer
- Interactive Reporting
- Dashboard Designer
- C-Tools

Client-Based Design Tools

- Data Integration
- Report Designer
- Metadata Editor
- Schema WorkBench

Enterprise Architecture

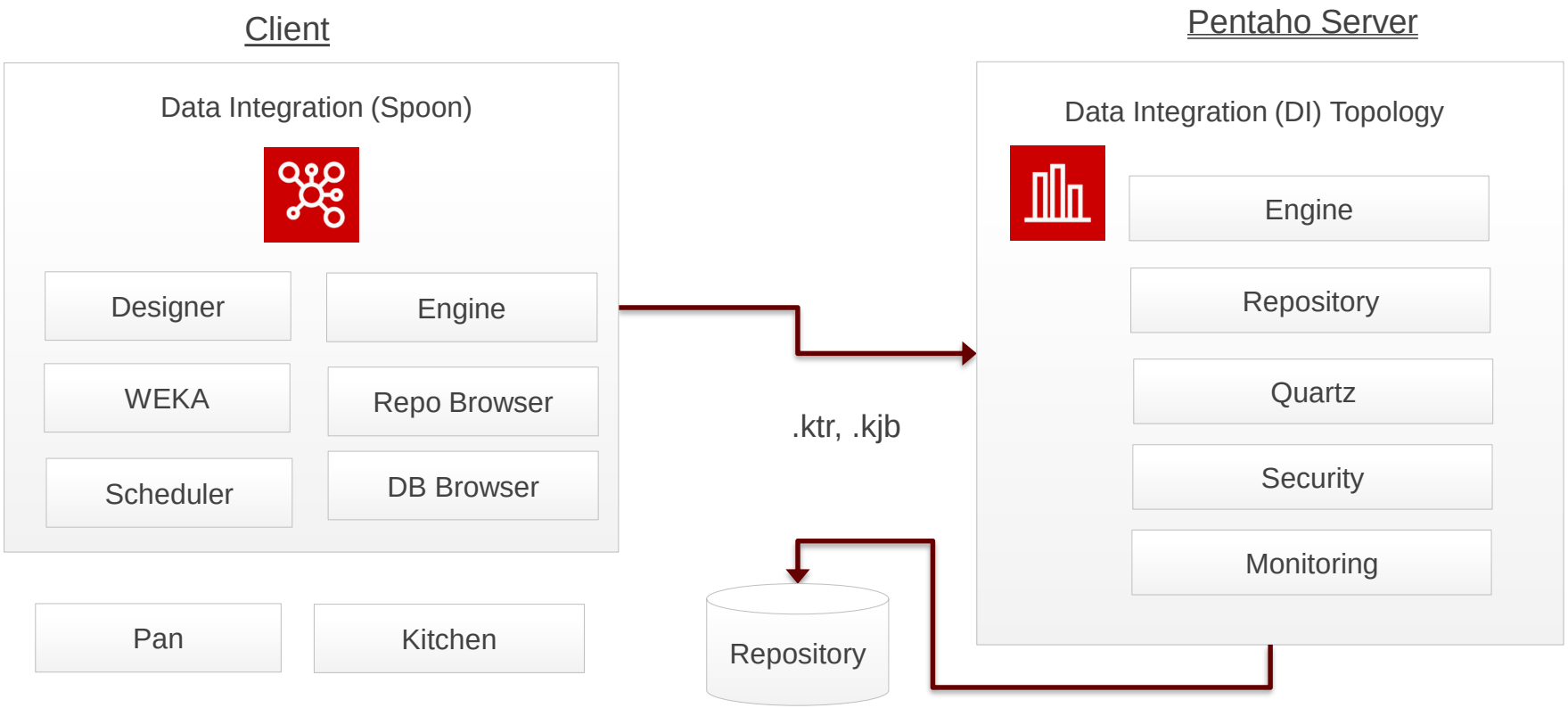
HITACHI
Inspire the Next



HITACHI
Inspire the Next

Components

Components



Components

HITACHI
Inspire the Next

Spoon

- Graphical modeling environment for developing, testing, debugging and monitoring jobs and transformations

Data Integration Server

- Dedicated ETL server used for remote execution providing scheduling, security integration and content management capabilities

Components

HITACHI
Inspire the Next

Carte

- Light-weight HTTP server used for remote execution and parallel execution of jobs and transformations on a scale-out cluster (slave nodes)

Kitchen, Pan

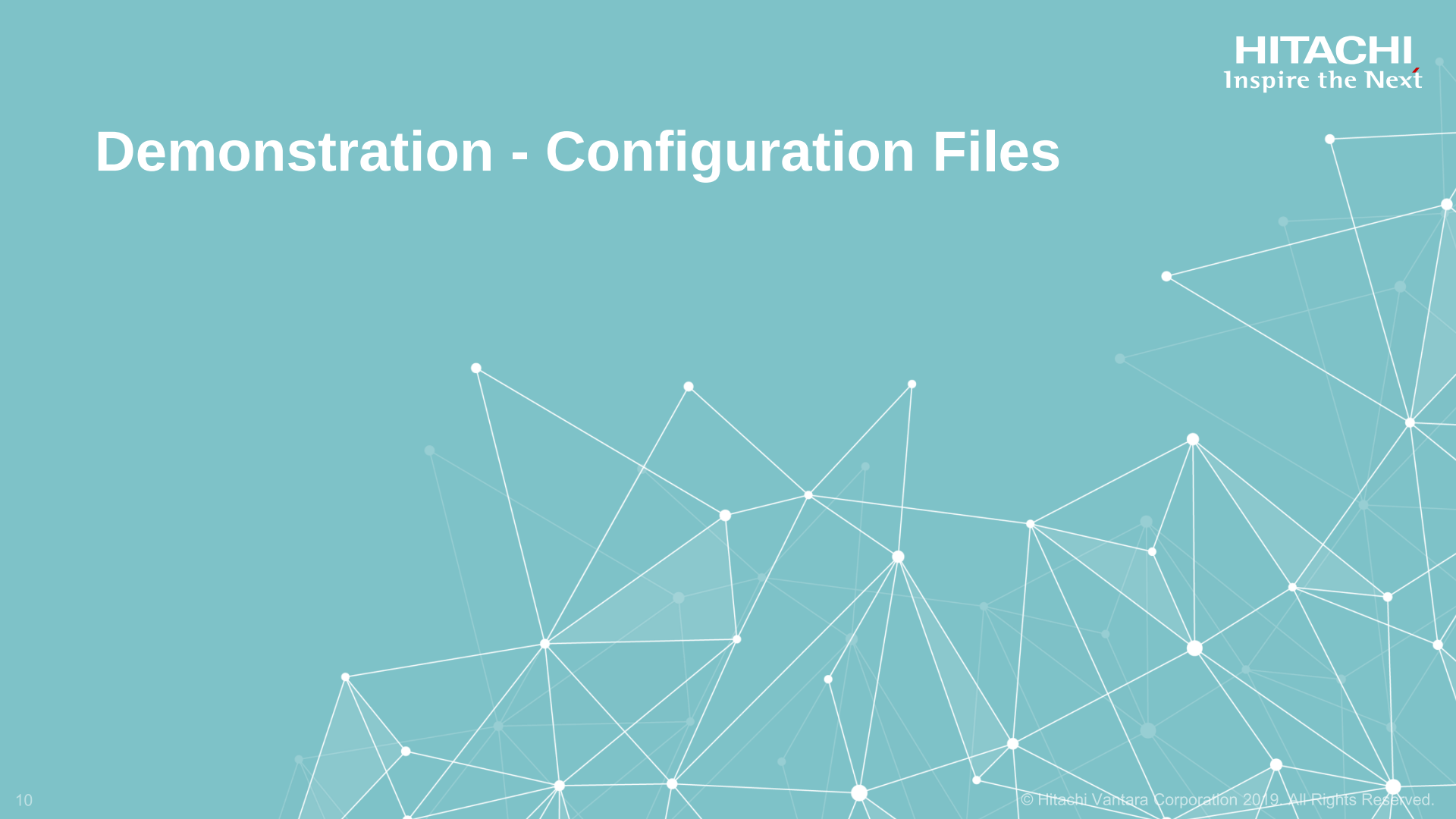
- Command-line driven job and transformation runners used for OS-level scheduling

HITACHI
Inspire the Next

Configuration

HITACHI
Inspire the Next

Demonstration - Configuration Files



Lab 1 - Configuration Directories

HITACHI
Inspire the Next

- All Kettle programs can be started using shell scripts located in the Kettle home directory.

Directory / File	Windows / Unix	Action
Shell script	spoon.bat / spoon.sh	Starts Spoon
Shell script	kichen.bat / kitchen.sh	Starts command line for Jobs
Shell script	pan.bat / pan.sh	Starts command line for Transformations
Samples	..\data-integration\samples	samples of .ktr and .ktj
..\lib	.jar files	add 'driver.jar' file
..\server\data-integration-server\tomcat\lib	.jar files	add 'driver.jar' file

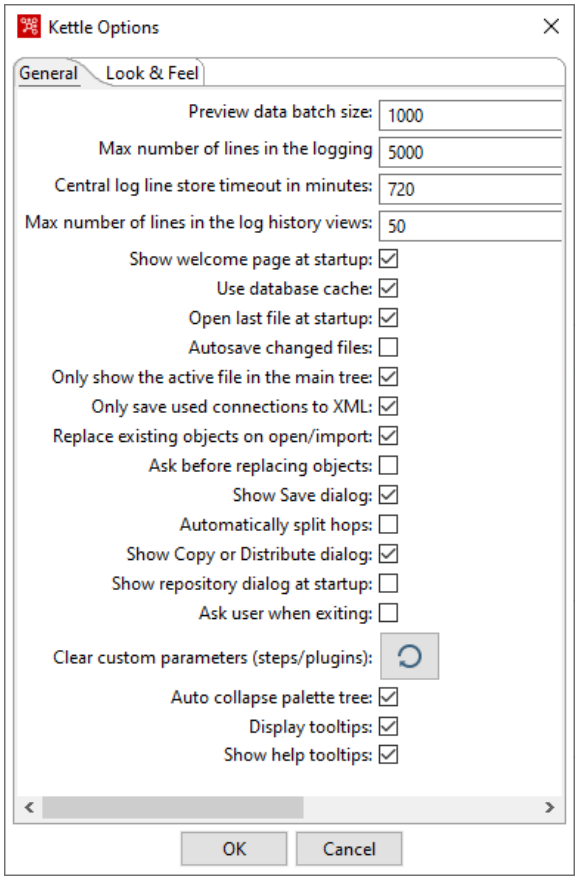
Lab 1 - kettle.properties

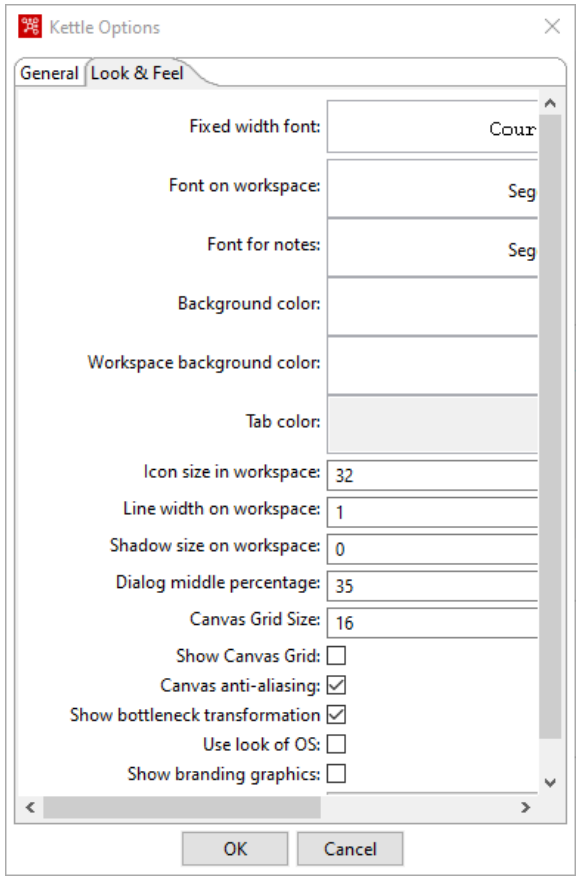
HITACHI
Inspire the Next

- Kettle home directory (.kettle)
  - Configuration files that control the behaviour of PDI jobs and transformations
  - Located at user's home directory by default user dependency
  - A variable can be used set the location of property files for projects

kettle.properties	Default properties file for variables
shared.xml	Default shared objects file
db.cache	The database cache for metadata
repositories.xml	The local repositories file
.spoonrc	User interface settings, last opened transformation/job
.languageChoice	User language (delete to revert language)

Lab 1 - UI Configuration





Lab 1 – kettle.properties

HITACHI
Inspire the Next

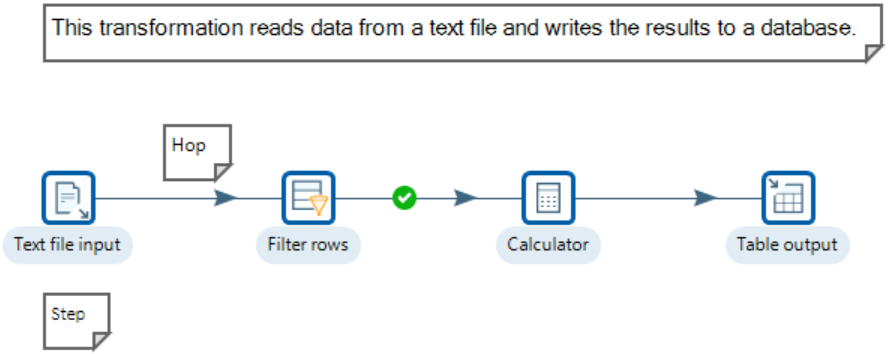
- Main configuration file
- Used to reference other required artefacts for ETL projects

```
## This file was generated by Pentaho Data Integration version 8.3.0.0-371.
#
# Here are a few examples of variables to set:
#
# PRODUCTION_SERVER = hercules
# TEST_SERVER = zeus
# DEVELOPMENT_SERVER = thor
#
# Note: lines like these with a # in front of it are comments
#
#

KETTLE_LOG_SCHEMA=pentaho_dilogs

KETTLE_CHANNEL_LOG_DB=live_logging_info
KETTLE_CHANNEL_LOG_SCHEMA=pentaho_dilogs
KETTLE_CHANNEL_LOG_TABLE=channel_logs

KETTLE_METRICS_LOG_DB=live_logging_info
KETTLE_METRICS_LOG_TABLE=metrics_logs
KETTLE_METRICS_LOG_SCHEMA=pentaho_dilogs

KETTLE_TRANS_LOG_DB=live_logging_info
KETTLE_TRANS_LOG_SCHEMA=pentaho_dilogs
KETTLE_TRANS_LOG_TABLE=trans_logs
```

HITACHI
Inspire the Next

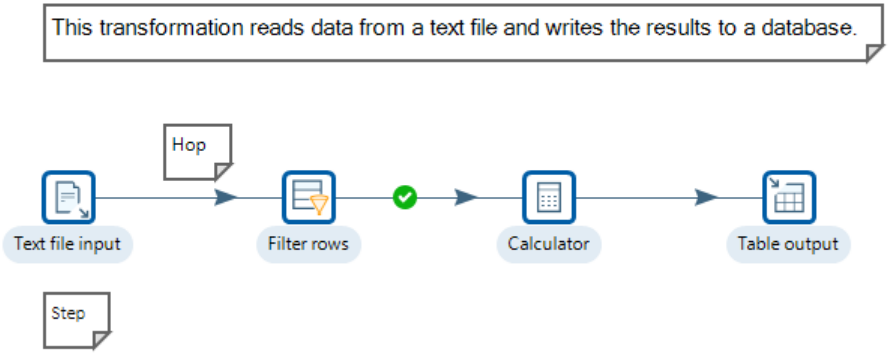
Transformations

Transformations



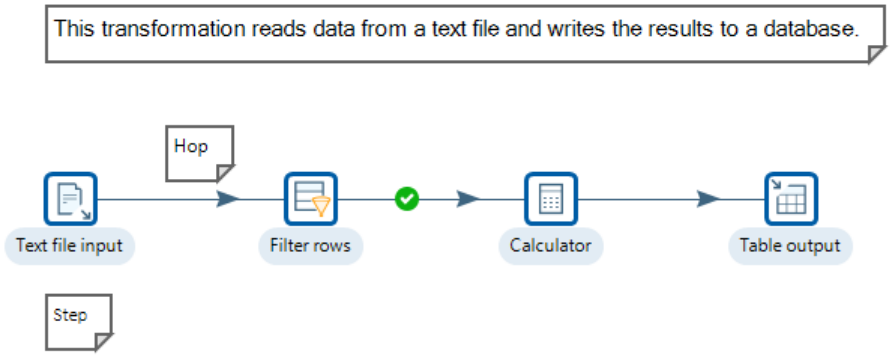
- Workhorses of your ETL solution
- Handle the manipulation of rows of data
- Consist of steps that perform the core work
- Steps are connected by hops
- Data Flow is the movement of rows from one step to another

Transformations



- Core building blocks in a transformation, each having their distinct functionality
- Read data from incoming hops, write data to outgoing hops
- Can have multiple outgoing hops: copy or distribute (round robin)
- Each step is started simultaneously and runs in its own thread - Parallelism
- A thread is a concurrently running task

Hops

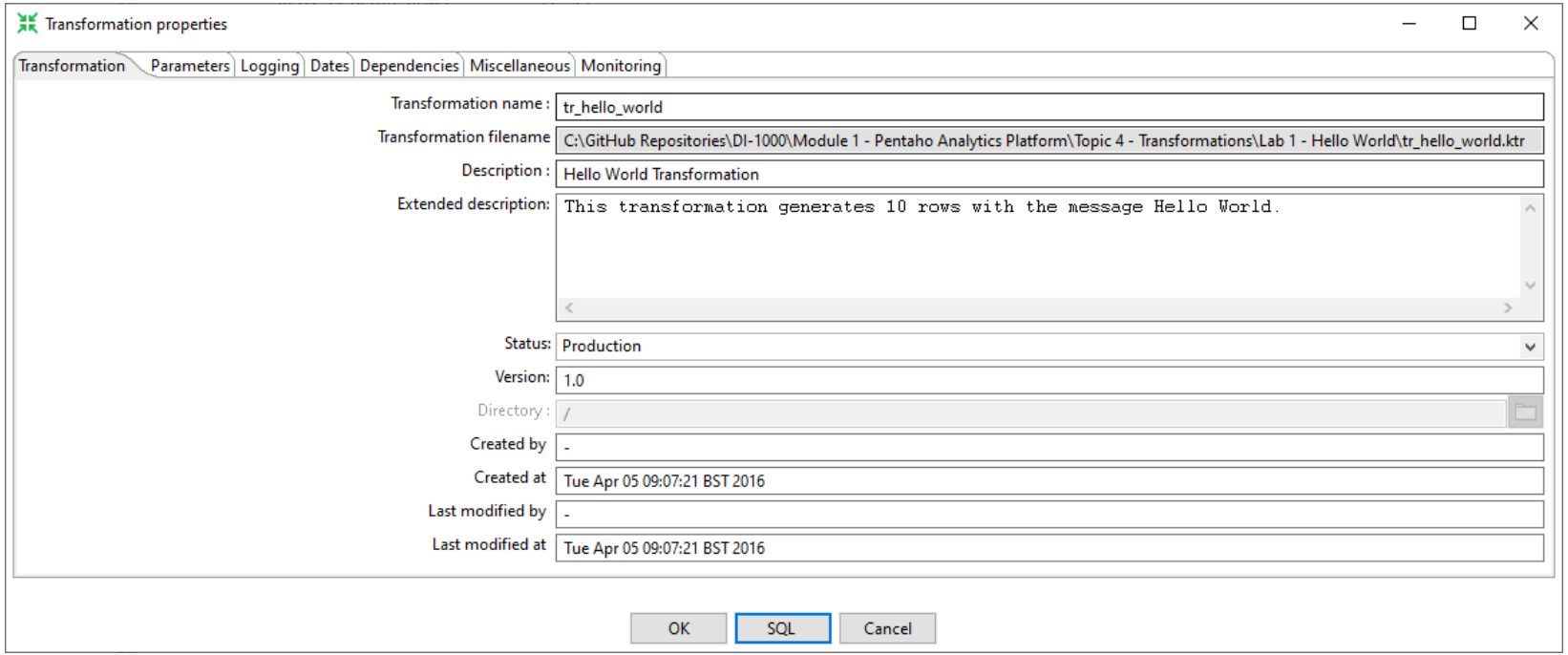


- Define the data path between steps
- Hops also represent a row buffer called the row set (5,000 – 50,000)
- When row set is full, the step that writes rows halts
- Different types of hops available

Transformation Properties

HITACHI
Inspire the Next

- Each Transformation has a set of associated properties



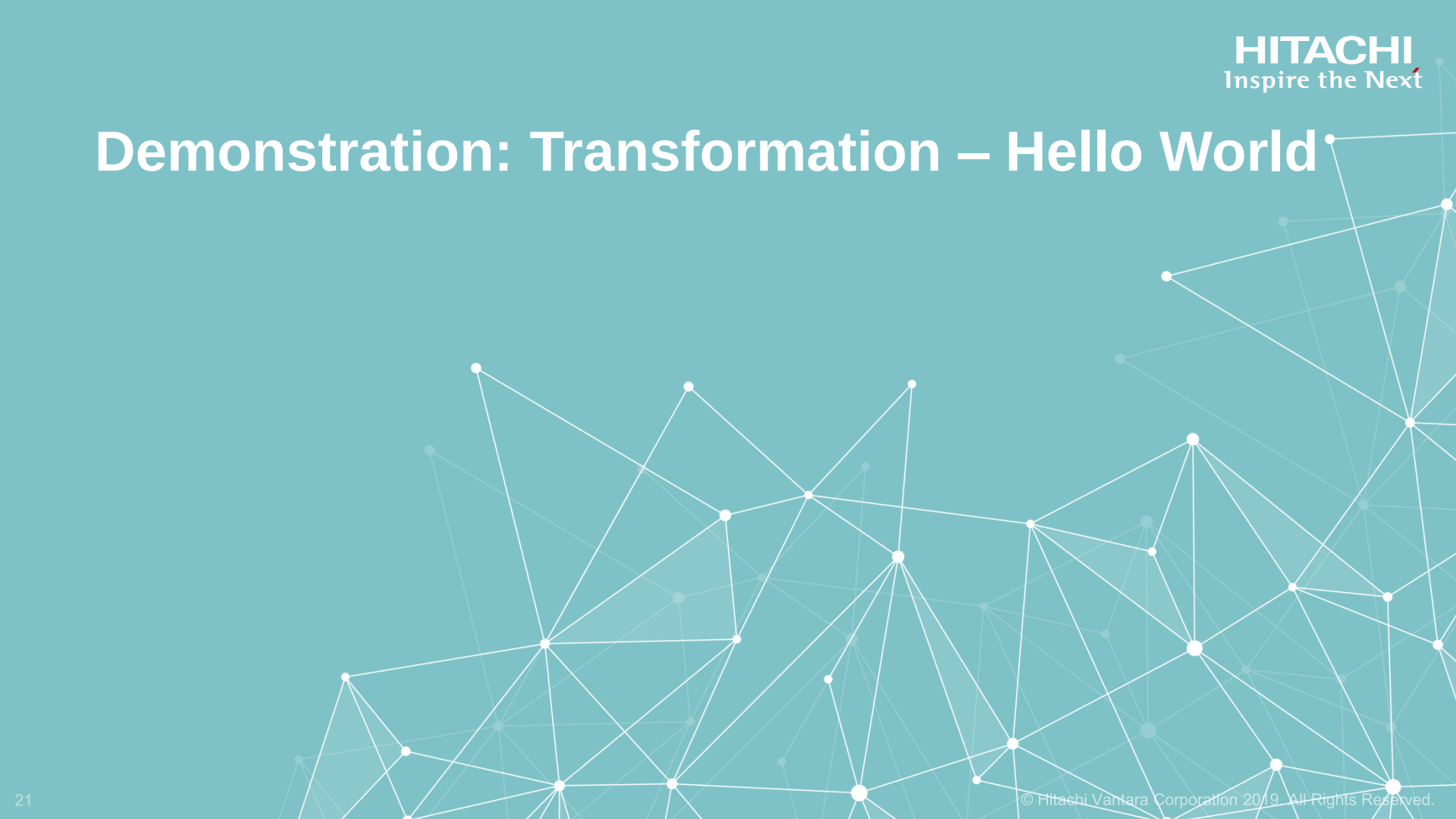
Parallelism

HITACHI
Inspire the Next

- The hops (buffer) allow steps to be executed in parallel
  - Work independently, at their own speed, in separate threads
- Once a buffer is full, parallelism gets reduced
  - Steps will operate at virtually the same speed
- Not possible to define an order of execution in a transformation
  - Every step is started simultaneously
  - Rows are being forced through the step network
- Functionally a transformation does have a start and end
- If you need to perform tasks in a specific order, refer to Jobs

HITACHI
Inspire the Next

Demonstration: Transformation – Hello World



Demonstration: Hello World

HITACHI
Inspire the Next

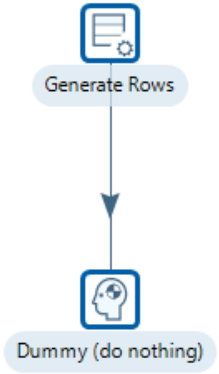
This transformation generates rows, 'hello world'. Illustrates some of the key features:

Step1: Generate Rows
Limit: 10 rows

Name: message
Type: String
Value: hello world

click: Preview

Step2: Dummy



- Generate ‘hello world’ data
- Explore execution results

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

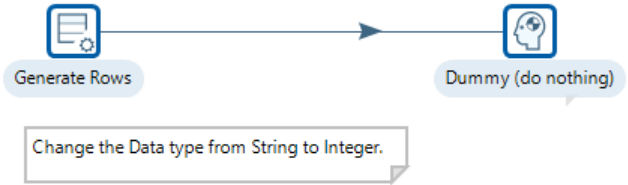
#	message
1	hello world
2	hello world
3	hello world
4	hello world
5	hello world

Demonstration: Hello World

HITACHI
Inspire the Next

- Change the data type in ‘hello world’ from String to Integer

Changing the format from String to Integer results in an error.



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

```
2019/09/04 10:29:01 - Spoon - Started the transformation execution.
2019/09/04 10:29:01 - tr_hello_world - Dispatching started for transformation [tr_hello_world]
2019/09/04 10:29:01 - Hello World.0 - ERROR (version 8.3.0.0-371, build 8.3.0.0-371 from 2019-06-11 11.09.08 by buildguy) : Couldn't parse Integer field [messa
2019/09/04 10:29:01 - Hello World.0 - Unexpected conversion error while converting value [message String] to an Integer
2019/09/04 10:29:01 - Hello World.0 -
2019/09/04 10:29:01 - Hello World.0 - message String : couldn't convert String to Integer
2019/09/04 10:29:01 - Hello World.0 -
2019/09/04 10:29:01 - Hello World.0 - message String : couldn't convert String to number : non-numeric character found at position 1 for value [hello world]
2019/09/04 10:29:01 - Hello World.0 - ERROR (version 8.3.0.0-371, build 8.3.0.0-371 from 2019-06-11 11.09.08 by buildguy) : Error initializing step [Hello World]
2019/09/04 10:29:01 - tr_hello_world - ERROR (version 8.3.0.0-371, build 8.3.0.0-371 from 2019-06-11 11.09.08 by buildguy) : Step [Hello World.0] failed to initial
2019/09/04 10:29:01 - Spoon - ERROR (version 8.3.0.0-371, build 8.3.0.0-371 from 2019-06-11 11.09.08 by buildguy) : tr_hello_world: preparing transformation e
2019/09/04 10:29:01 - Spoon - ERROR (version 8.3.0.0-371, build 8.3.0.0-371 from 2019-06-11 11.09.08 by buildguy) : org.pentaho.di.core.exception.KettleExcep
2019/09/04 10:29:01 - Spoon - We failed to initialize at least one step.  Execution can not begin!
```

Module Review

HITACHI
Inspire the Next

In this module, you should have learned to:

- Outline the Pentaho Platform
- Configure & Describe key Pentaho Data Integration components
- Understand the concept of Parallelism
- Define Transformations

HITACHI
Inspire the Next

Thank You

HITACHI

Inspire the Next