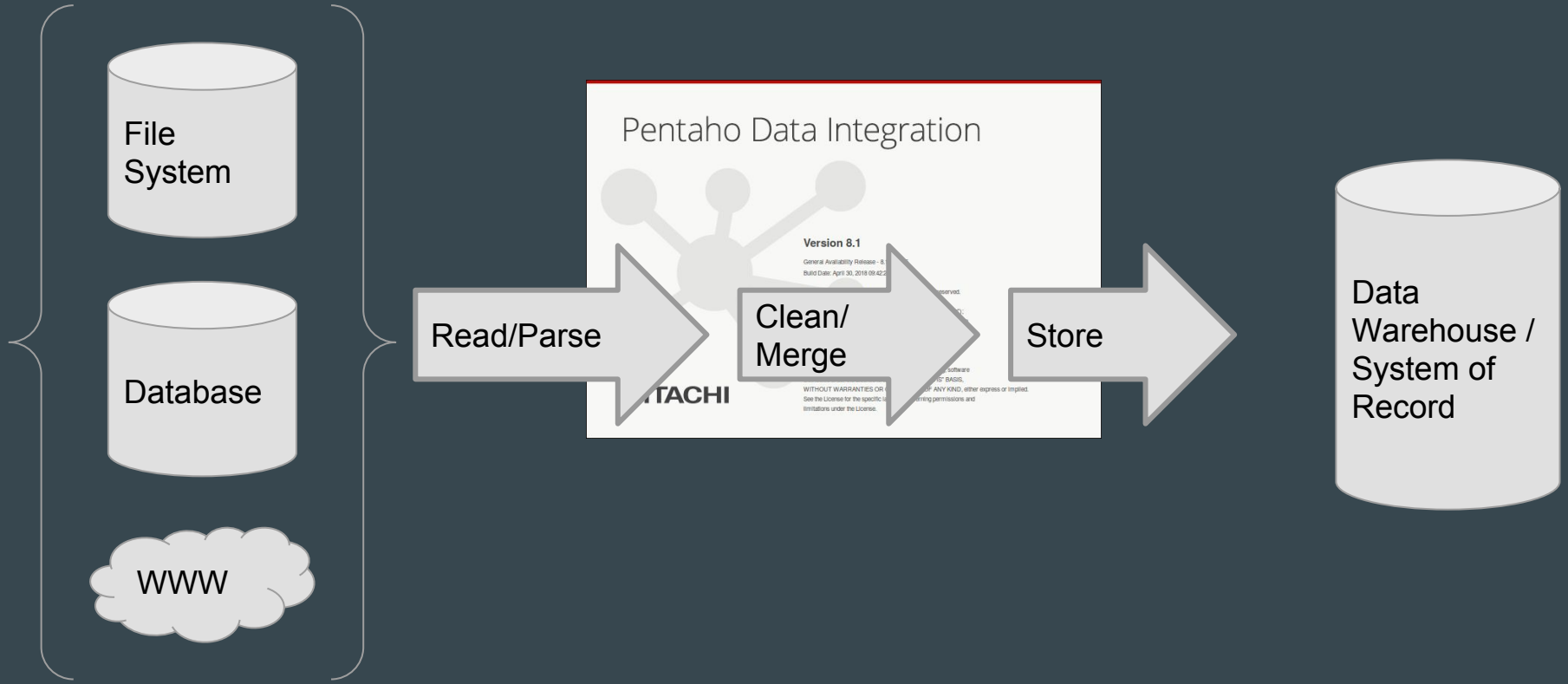
Introduction to Pentaho Kettle

Nick Hudak
nhudak3@gmail.com

Topics

- What is ETL?
- Kettle's capabilites
- Installation
- Navigating Kettle
- Transformations
- Data Sources
- Jobs

ETL: Extract ⇒ Transform ⇒ Load



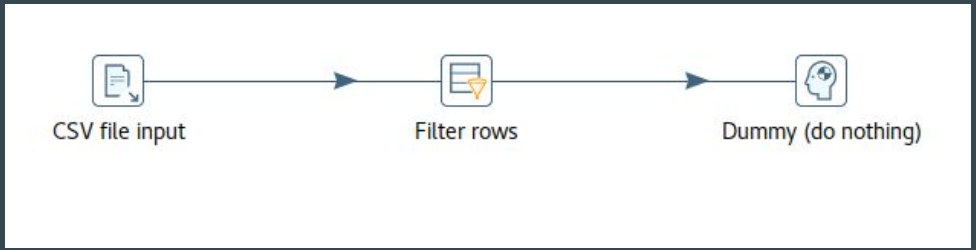
What can Kettle do?

- Graphical ETL designer simplifies the creation of data pipelines
- Rich library of prebuilt components help to access, prepare and blend data
- Powerful orchestration capabilities coordinate and combine transformations

Transformation or Job?

Transformation:

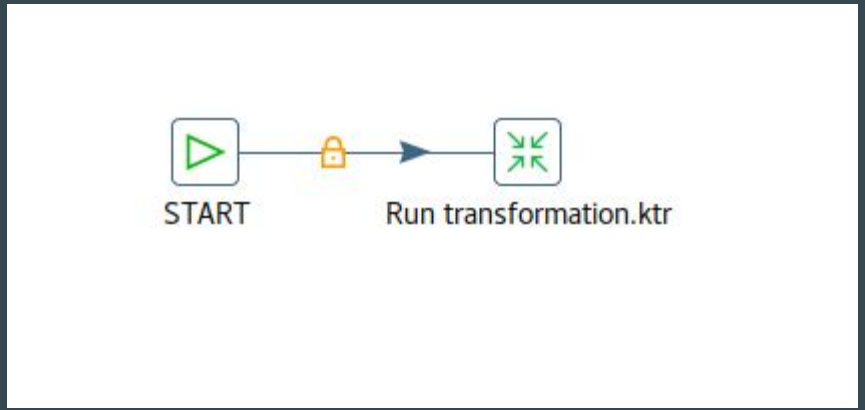
- Think “Pipeline”
- All transformation steps run at the same time
- Processes tabular data, rows and columns
- Starts at every input step
- Rows pass from one step to the next over
- All branches followed simultaneously
- Saved as *.ktr* file



Transformation or Job?

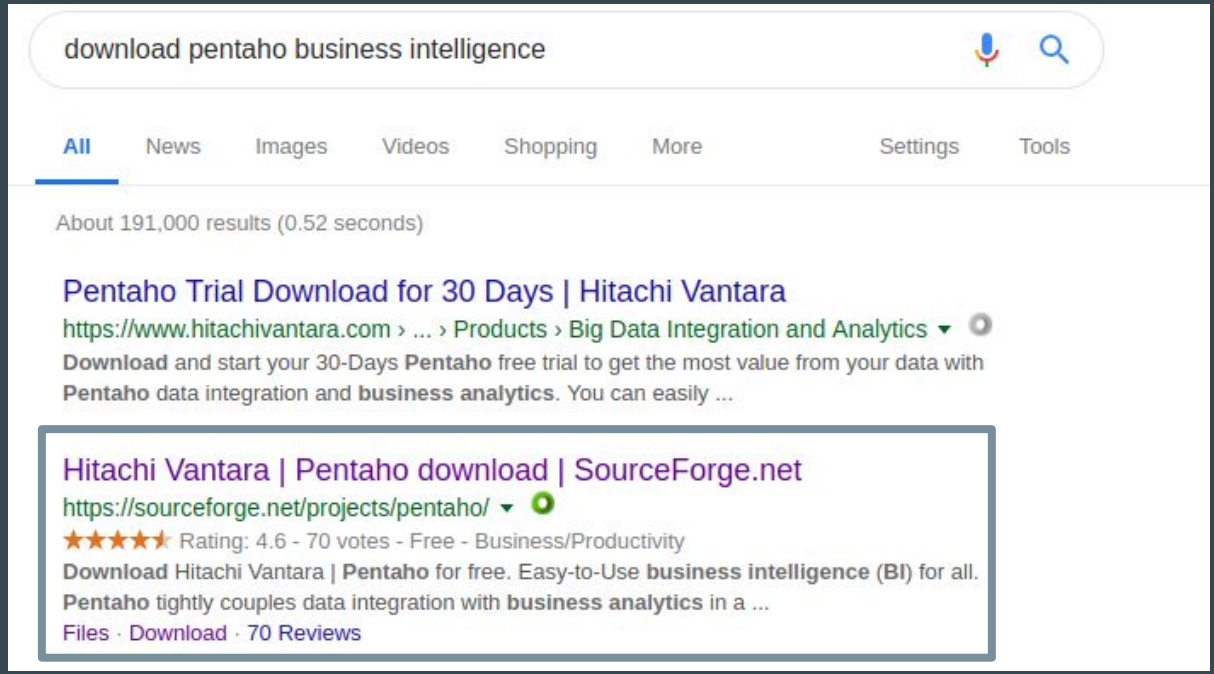
Job

- Think “Flowchart”
- Job entries execute one after another
- Single starting point
- Only one path executes
- Some entries can split down different paths
- Saved as *.kjb* file



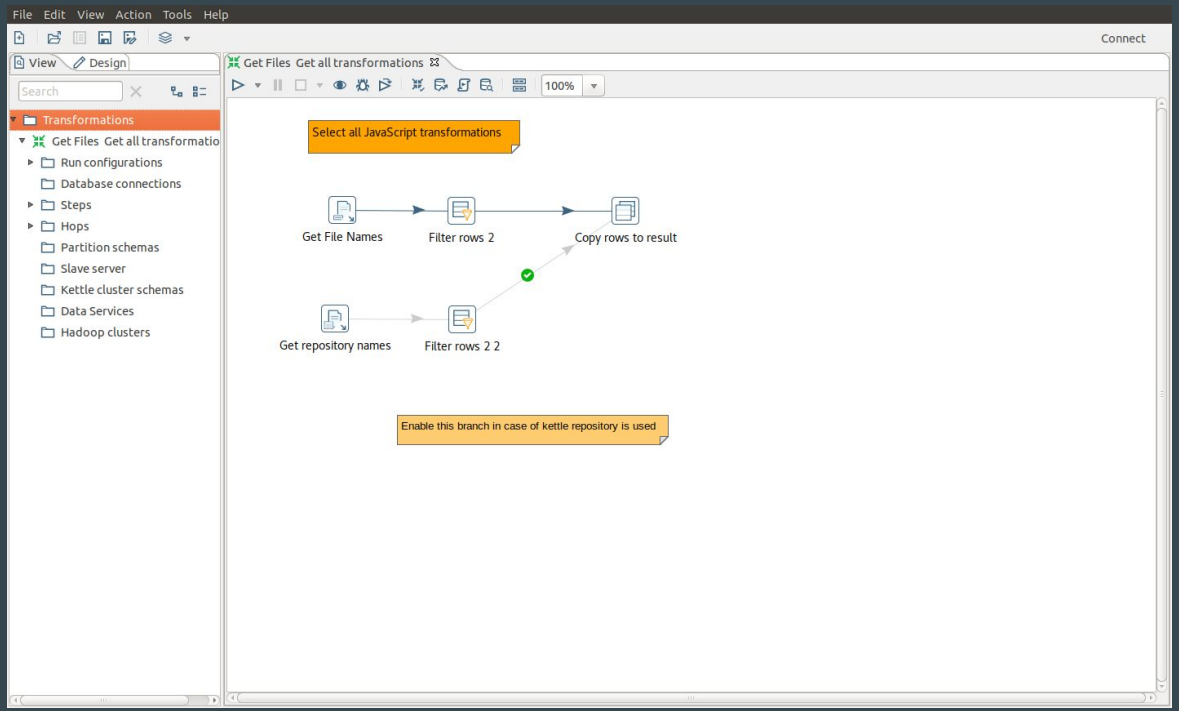
Installation

- Search for some of:
  - Hitachi Vantara
  - Pentaho
  - Kettle
  - Data Integration
  - Business Intelligence
  - (they're all the same!)
- Download latest version from SourceForge
- Don't need the server
- Community Edition (ce)
  - No trial, free forever
- Unpack and run data-integration/Spoon.bat



Exploring Spoon

http://help.pentaho.com ⇒ Products ⇒ Data Integration



Reading Local Data

- CSV
  - Easy import/export to Excel
  - Specify column types
- XML
  - XPath - See examples at W3Schools
  - Extract elements at any depth
  - Additional fields to add file name, date, etc...
- JSON
  - Similar function to XML input
  - JsonPath - Read The Doc!

Reading and Storing from Databases

- Shared database connections
  - Specify vendor, host/port, username and password
  - Same connection can be used by many steps
  - Use “share” to reuse in other transformations
  - “Explore” to preview database structure
- Table Output step
  - Assumes transformation fields have same names and types as table
  - Can be manually specified
  - “SQL” button can setup table for you
- Table Input step
  - Read rows from arbitrary SQL statements