DATA IS IN OUR DNA

Guided Analytics at Seagate

Seagate Technology Operations & Technology Advanced Analytics Group

Allan Luk, Analytics Business Solutions Director
Dr. Eric Lin, Sr Staff Analytics Business Solutions Engineer

November 9, 2018

© Seagate Technology, LLC, 2018

SEAGATE

You May Know Seagate as a Hard Drive Manufacturer…

- #1 OEM storage
- 1st to build and ship a collective zettabyte to the world
- Stores more than 40% of the world's data

© Seagate Technology, LLC, 2018

But We're also a Company that:

Serves many types of customers and businesses

Delivers deep expertise and unique IP in storage & data management

Combines UX, software & design capabilities to create new categories of storage solutions

Ranks as one of the top 25 companies worldwide in supply chain operations

© Seagate Technology, LLC, 2018

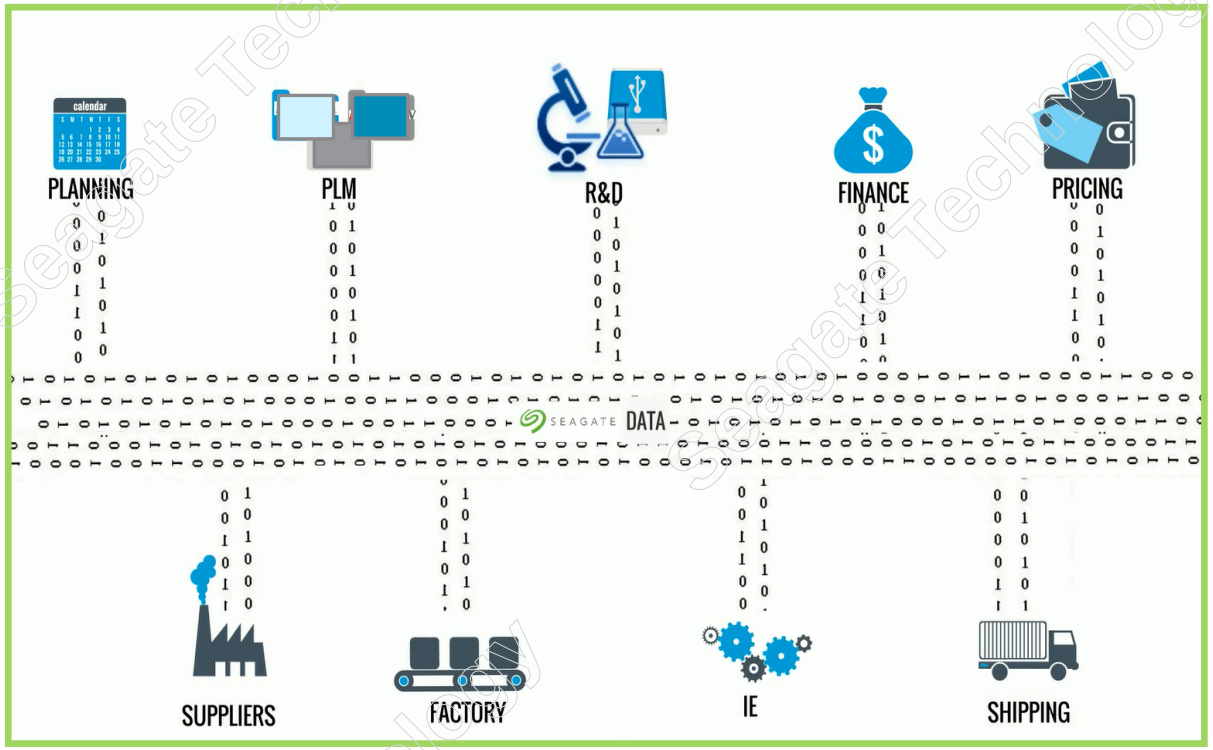
Presentation Overview

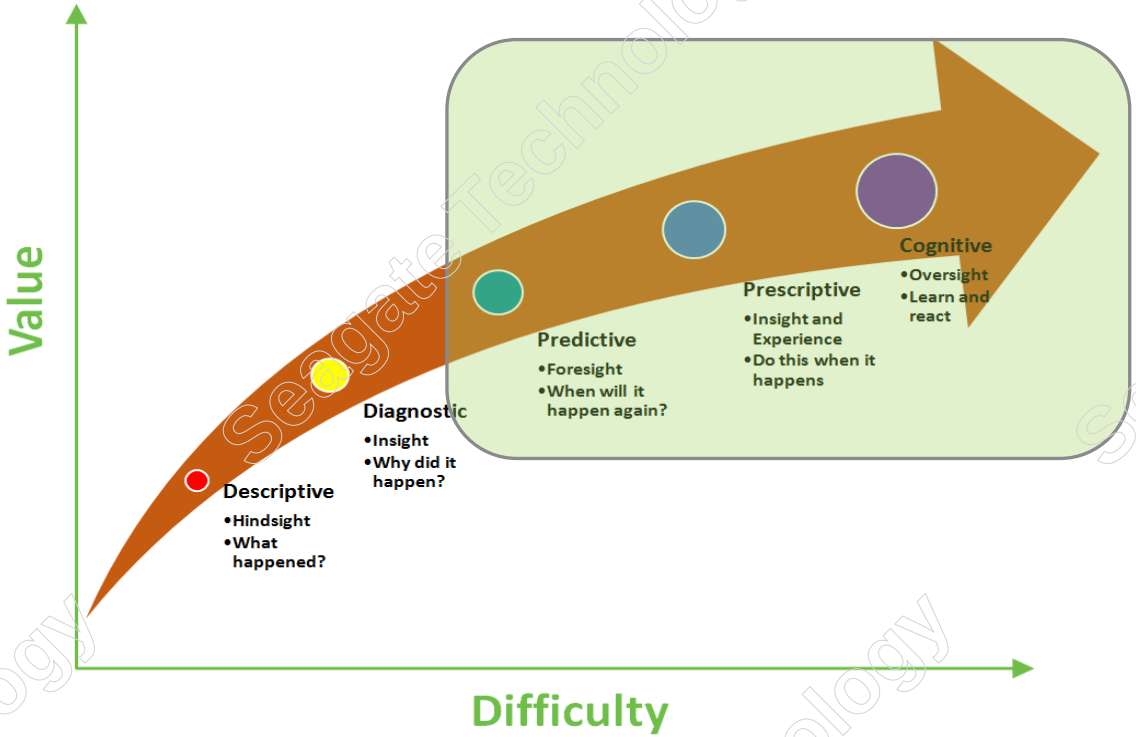
- **Guided Analytics at Seagate Update**
- **Guided Analytics Software Development and Deployment Challenges**
- **KNIME Development and Deployment Examples**
  - This Presentation (Allan Luk, Eric Lin) - Manufacturing, HR
  - Next Presentation (Debin Wang) - R&D, Engineering
- **Summary**
- **Live Demo:**
  - Parallel Execution of Data Source Query in KNIME

Seagate Technology

© Seagate Technology, LLC, 2018

Background





Tons of Data everywhere

Move up analytics capability curve

Analytics → INSIGHTS

Background

Citizen Data Science (CDS) Initiative @ Seagate
Turn Data into INSIGHTS

- CDS & Guided Analytics Software Workshop
- Guided Analytics Software Development & Deployment
- CDS Certification
- CDS Community Building

Open for Innovation® KNIME

Awareness → **Software & Tools** → **Capability** → **Ecosystem: Community Contributions**

Overview: Guided Analytics at Seagate

© Seagate Technology, LLC, 2018

Guided Analytics Software (KNIME) Workshop/Tech Talk

Flow Variables & Looping

- HGA data as an example to build correct Wafermaps
- SQL looping with chunks of batches per your choice

1

Database Connection

- Oracle
  - ODS
  - EDW
- Hadoop
  - EHC
  - Big Search
- F3 Log Service
- Google sheets

2

How to read difficult text file format through KNIME

- F3 Log parsing example

3

Data Horizontalization

- Essential data prep steps before building machine learning models

4

Model Optimization

- When models are complex with many parameters available for tuning, how do you optimize them?

5

KNIME Tech Talk with Eric 2018

Linear Regression

- How to build a Linear Regression model when many of your input variables are non-numeric
- Next steps when Linear Regression model fails

6

Multi-Spec analysis KNIME method

- How to build a KNIME workflow to conduct a multi-spec analysis

7

Tool Integration

- JMP integration
- Matlab integration
- Python integration
- R integration
- HTML integration

8

Workflow Streaming

- How to optimize your workflow throughput efficiency

9

Visualization Improvement / Alternatives

- d3 Javascript development
- JMP integration
- Python Plot
- R Plot
- Matlab integration
- JFreeChart

10

© Seagate Technology, LLC, 2018

Seagate Guided Analytics Journey

- **2nd year on this journey**

- **The numbers:**
  - **Exposed KNIME to over 700 employees within Seagate**
  - **About 70 KNIME users**
  - **What are we using KNIME for:**
    - 65% - ETL, getting data, data preparation and as an integration platform
    - 35% - Modeling and prediction

- **Avid KNIME users:**
  - **1st batch: Manufacturing, Engineering, R&D**
  - **2nd batch: Other functions and businesses (e.g. HR)**

- **Using KNIME Desktop version**

- **KNIME Server version**
  - **Acquired license**
  - **To integrate with new IT infrastructure**

© Seagate Technology, LLC, 2018

Guided Analytics Software Development and Deployment Challenges

- **Deploying to the whole company**
  - **To get familiar with KNIME**
    - **Make it easier for employees to learn and apply at workplace**

- **Comparison to the existing software and solutions**
  - **Legacy system**
    - Ability to integrate with existing tools and software
  - **Features. Ease of use.**

- **Customerization for Seagate use cases:**
  - **Why customerization?**
    - **User experience**
    - **Connection to various data sources**
    - **Integration to the new IT infrastructure**
      - Analytics performed at the data source

- **Our wish list about the software:**
  - **Visualization and dashboard**
    - Features. Ease of use.
  - **Data exploration – more interactive**

© Seagate Technology, LLC, 2018

KNIME Development Examples

Objectives:

i) Improve Efficiency,

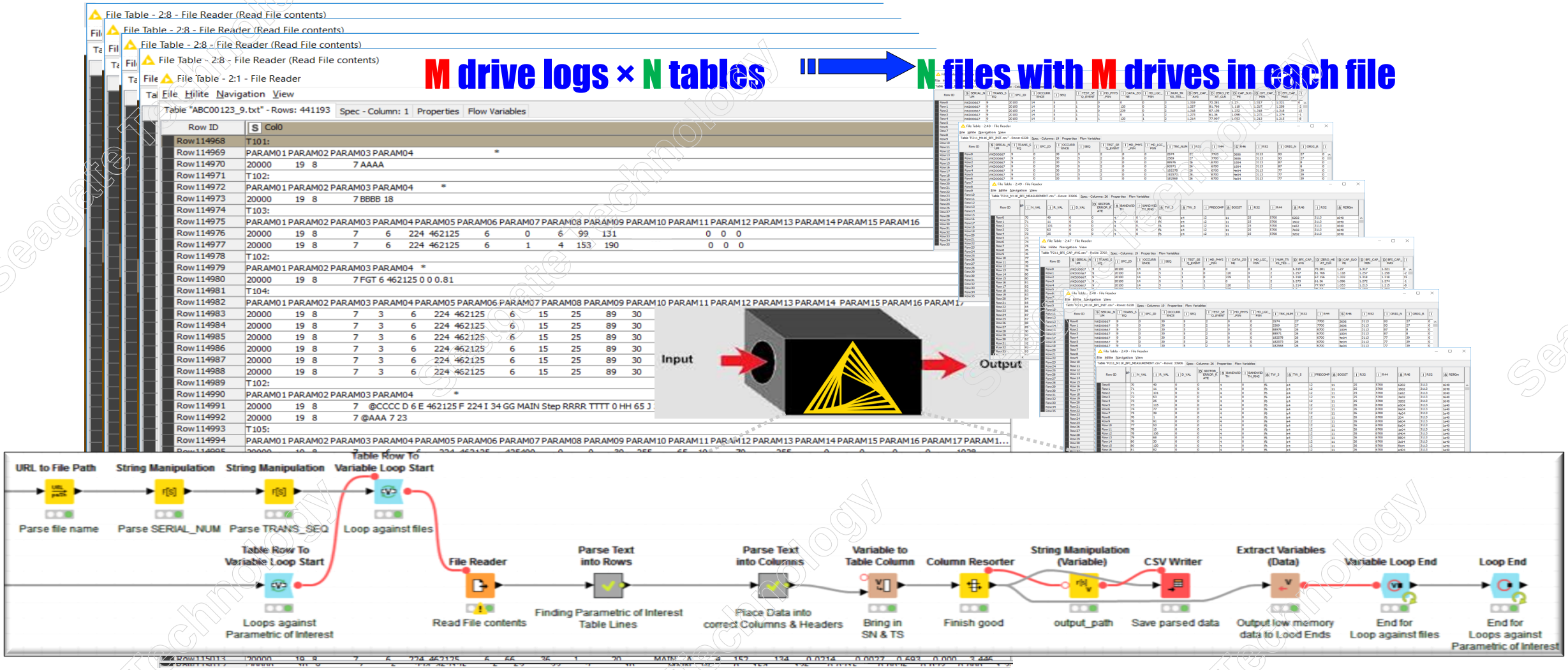
ii) Enable Collaboration and

iii) Enhance Adoption

© Seagate Technology, LLC, 2018

KNIME Development Example 1 – Manufacturing Data Query Node

M drive logs × N tables → N files with M drives in each file



Enable Data Wrangling and Automation. Essential to Product Performance Analysis.

© Seagate Technology, LLC, 2018

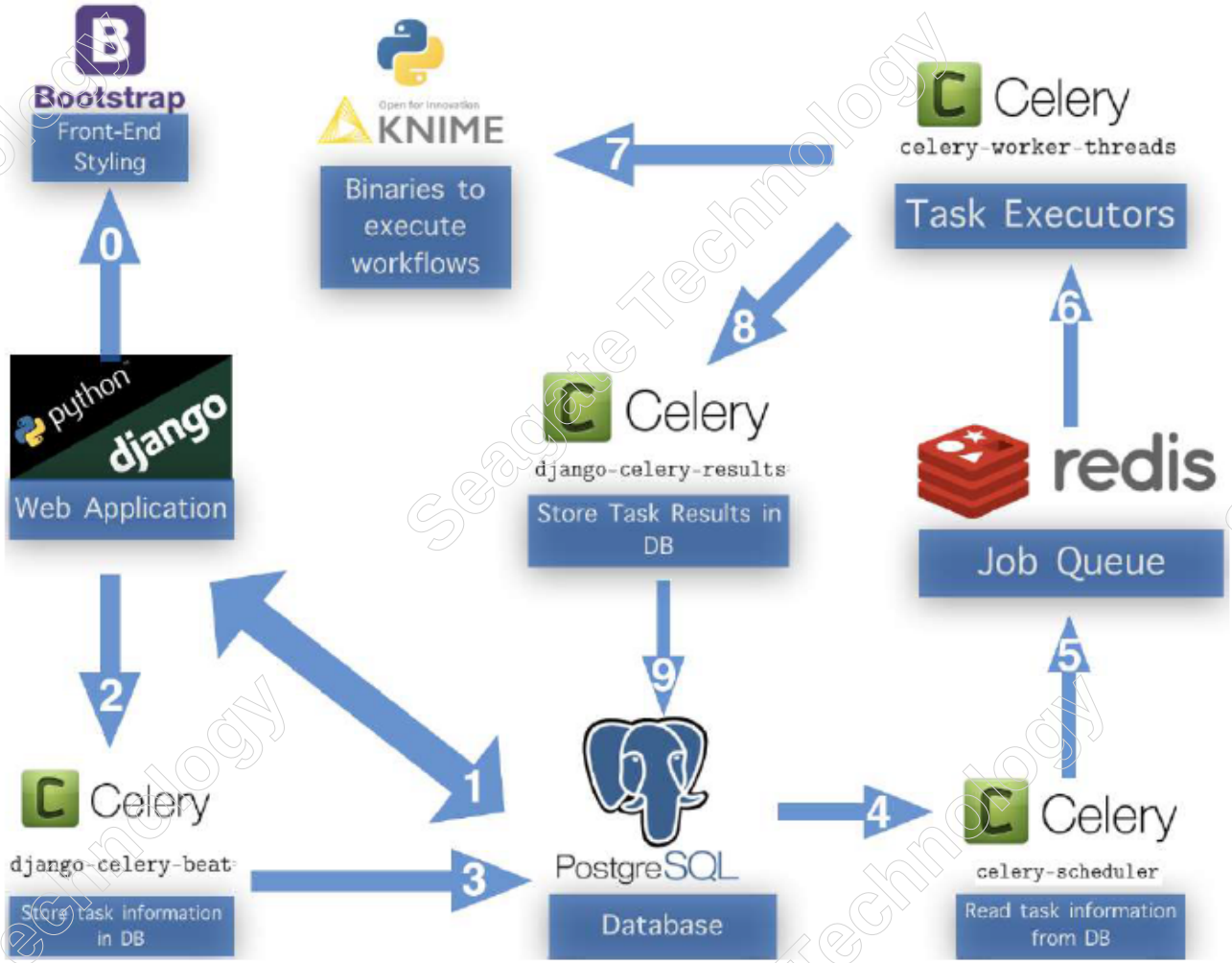
KNIME Development Example 2 – KNIME Containerization

Objectives:

- End-to-end Task Automation Suite for scheduling and automating various Data Science Workflows.
- Development of In-House Suite of Analytics Services.
- Containerization of KNIME workflows.

Deliverables:

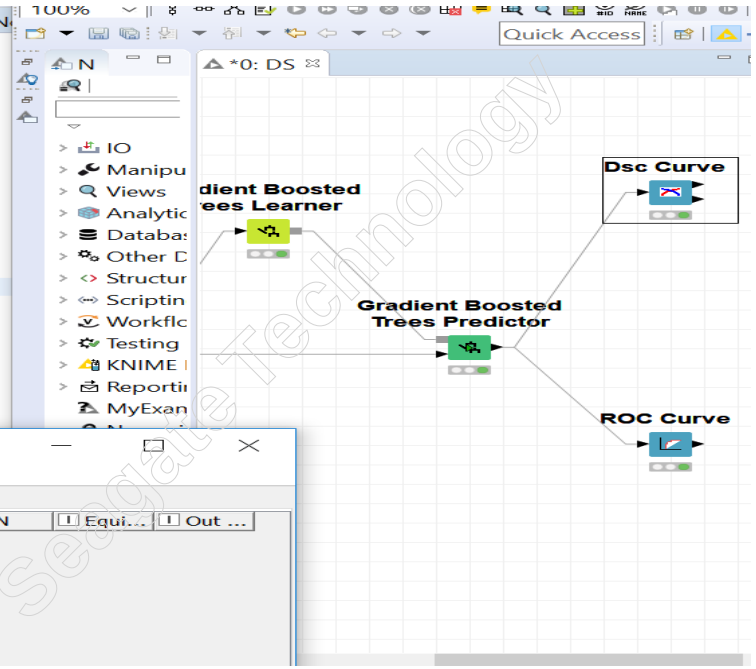
- A web application. Users upload data science workflow files.
- Users specify a schedule to run the workflow automatically.
- After the workflow runs, a report will be mailed to the users.

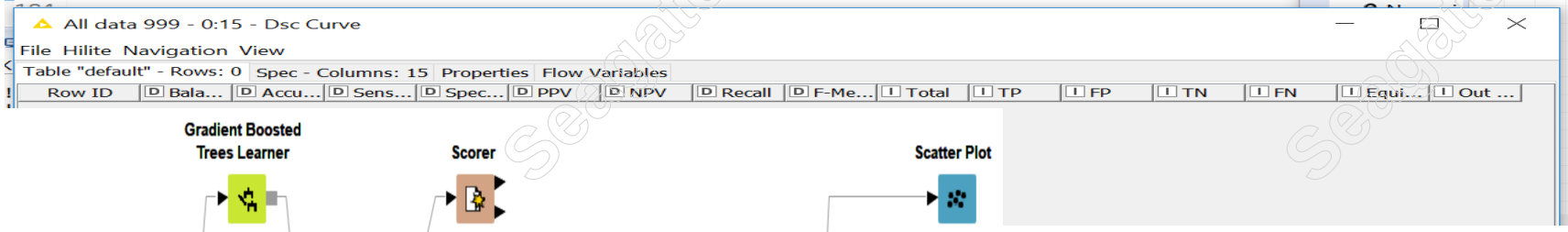


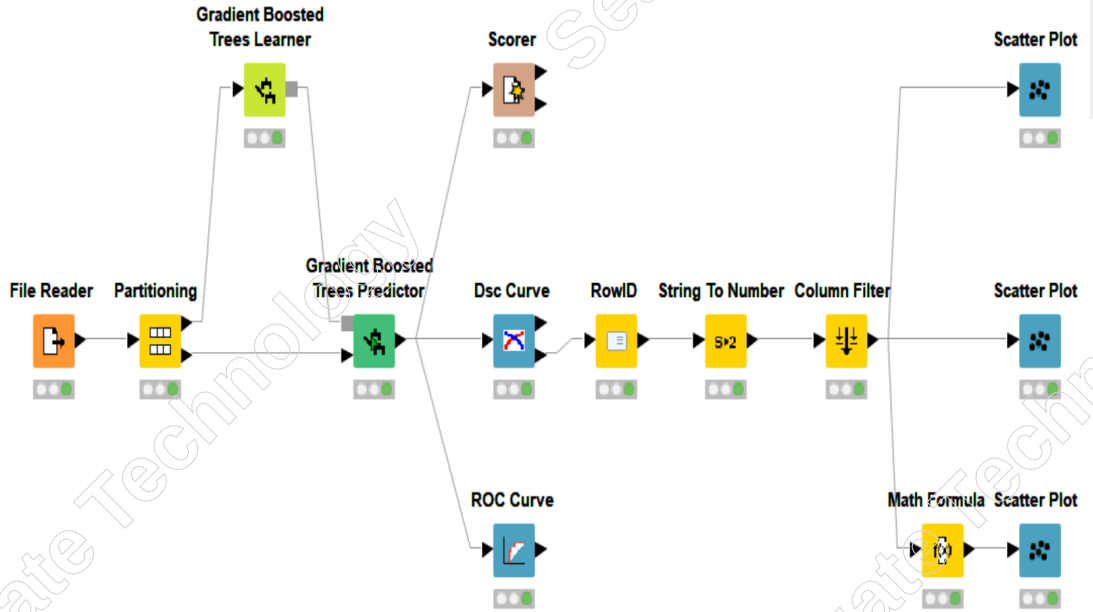
© Seagate Technology, LLC, 2018

KNIME Development Example 3 – Custom Node Creation

```
DscNodeMode...  DscSettings...  DscCalculato...  DscNodeView...  DscNodeFacto...  NodeFactory...
        if (calc.getWarningMessage() != null) {
            setWarningMessage(calc.getWarningMessage());
        }

        m_curves = calc.getOutputCurves();


        // Add one more table to show data
        DataTableSpec outputSpec = getOutputDataTableSpec();
        BufferedDataContainer container = exec.createDataContainer(outputSpec);

        /*for (int i = 0; i < results.size(); i++)
        {
            container.addRowToTable(createRow(new RowKey(binRanges[i]), results.get(i)));
        }
        */
        container.close();

        return new BufferedDataTable[] { calc.getOutputTable(), container.getTable() };
        //return new BufferedDataTable[] { calc.getOutputTable(), calc.getOutputTable() };

    }
```





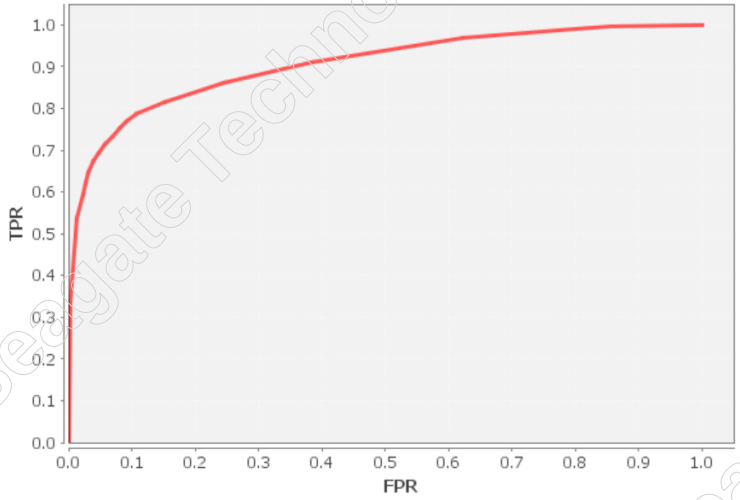


- **To augment ROC curve with additional all-in-one plot feature.**

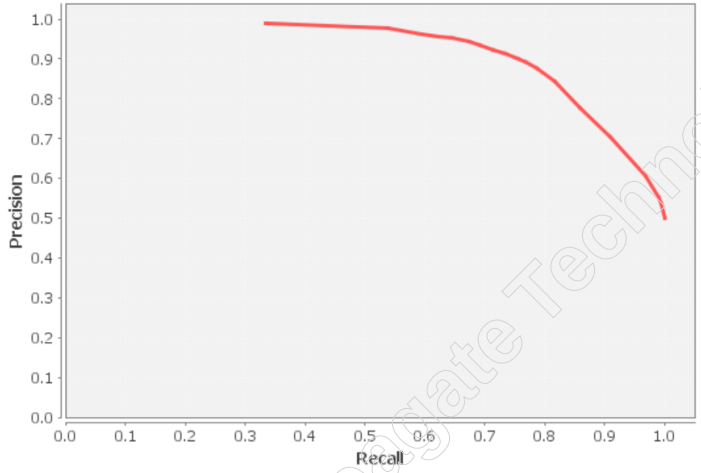
© Seagate Technology, LLC, 2018

KNIME Development Example 3 – Custom Node Creation

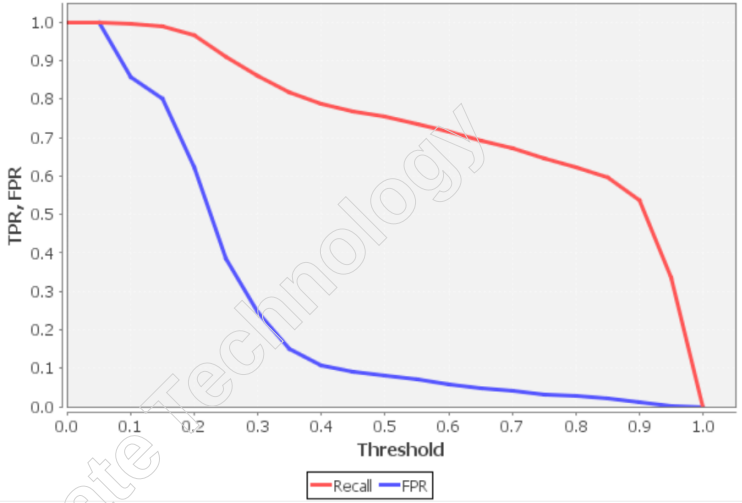
- **ROC: TPR (true positive rate) vs FPR (false positive rate)**



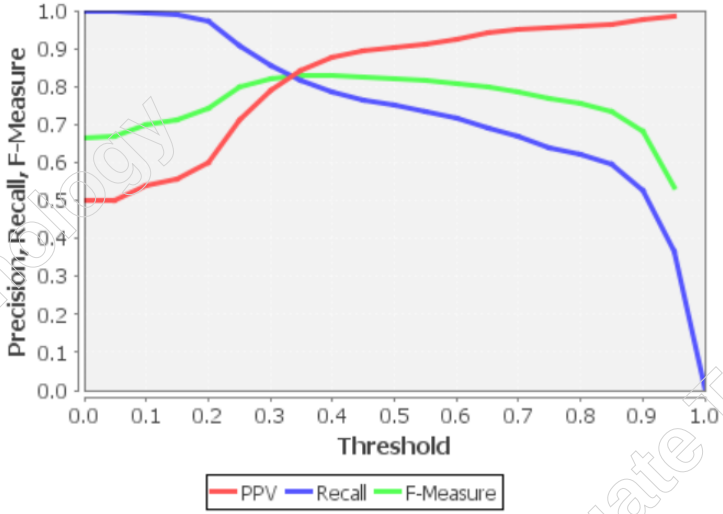
- **PRC: Precision vs Recall**



- **TPR, FPR vs Threshold**



- **F-Measure vs Threshold**



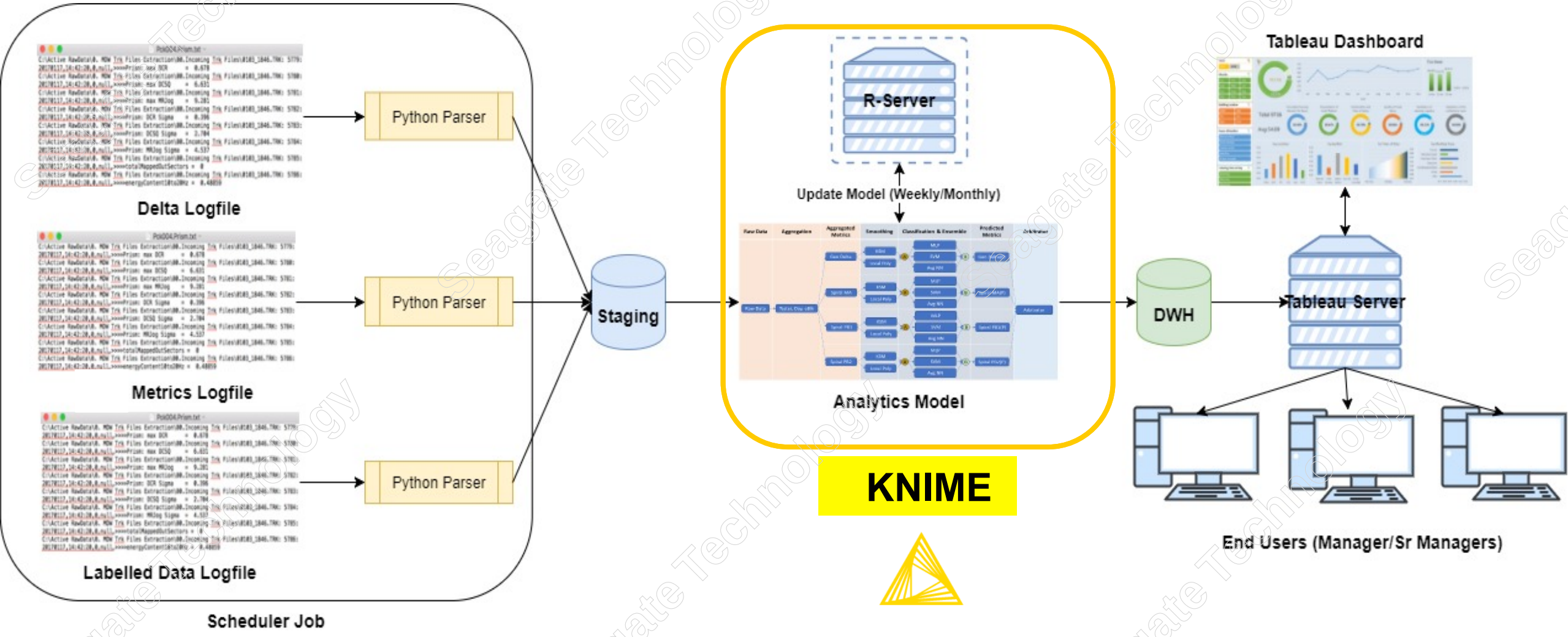
© Seagate Technology, LLC, 2018

KNIME Deployment Examples

i) Product Quality Assessment
ii) Equipment Health Prediction
iii) Manufacturing Image Analytics
iv) HR Analytics

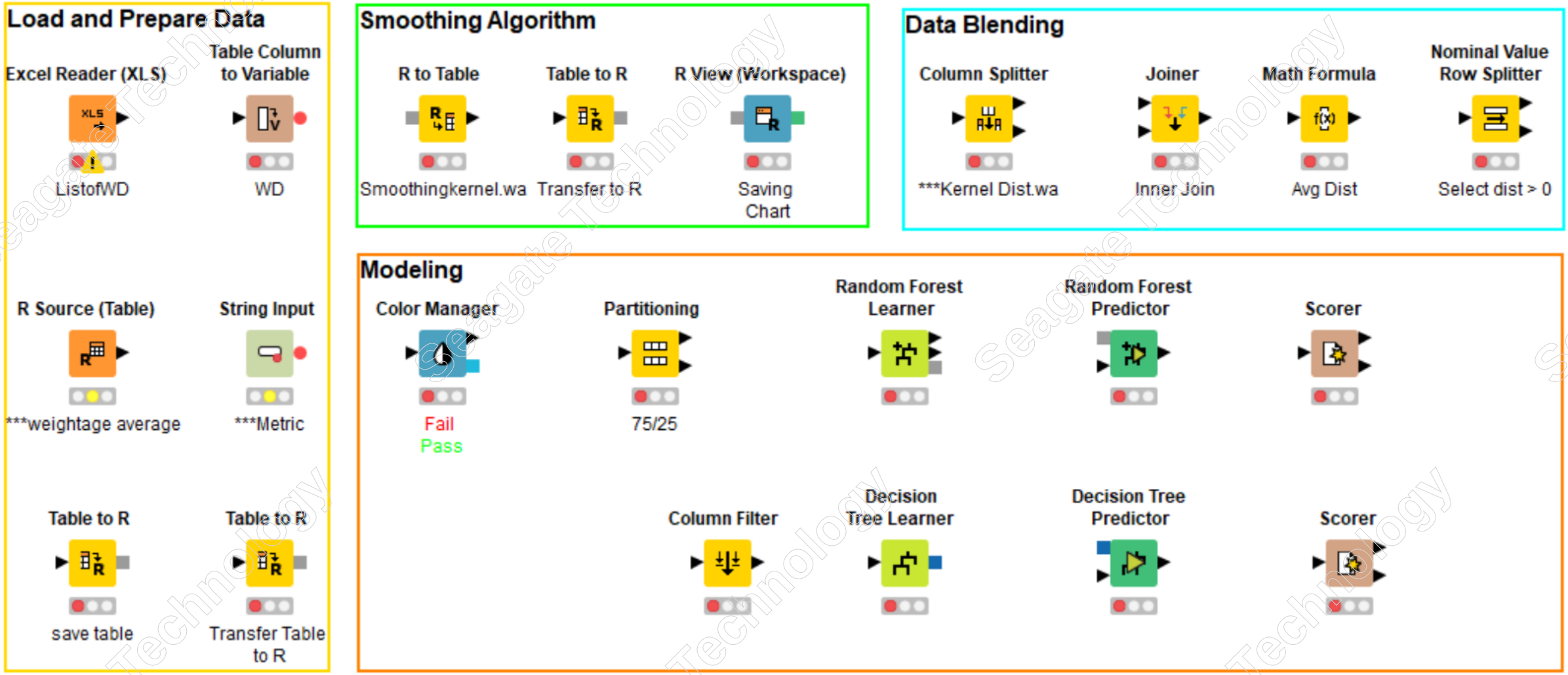
© Seagate Technology, LLC, 2018

KNIME Deployment Example 1 – Component Manufacturing Quality Assessment



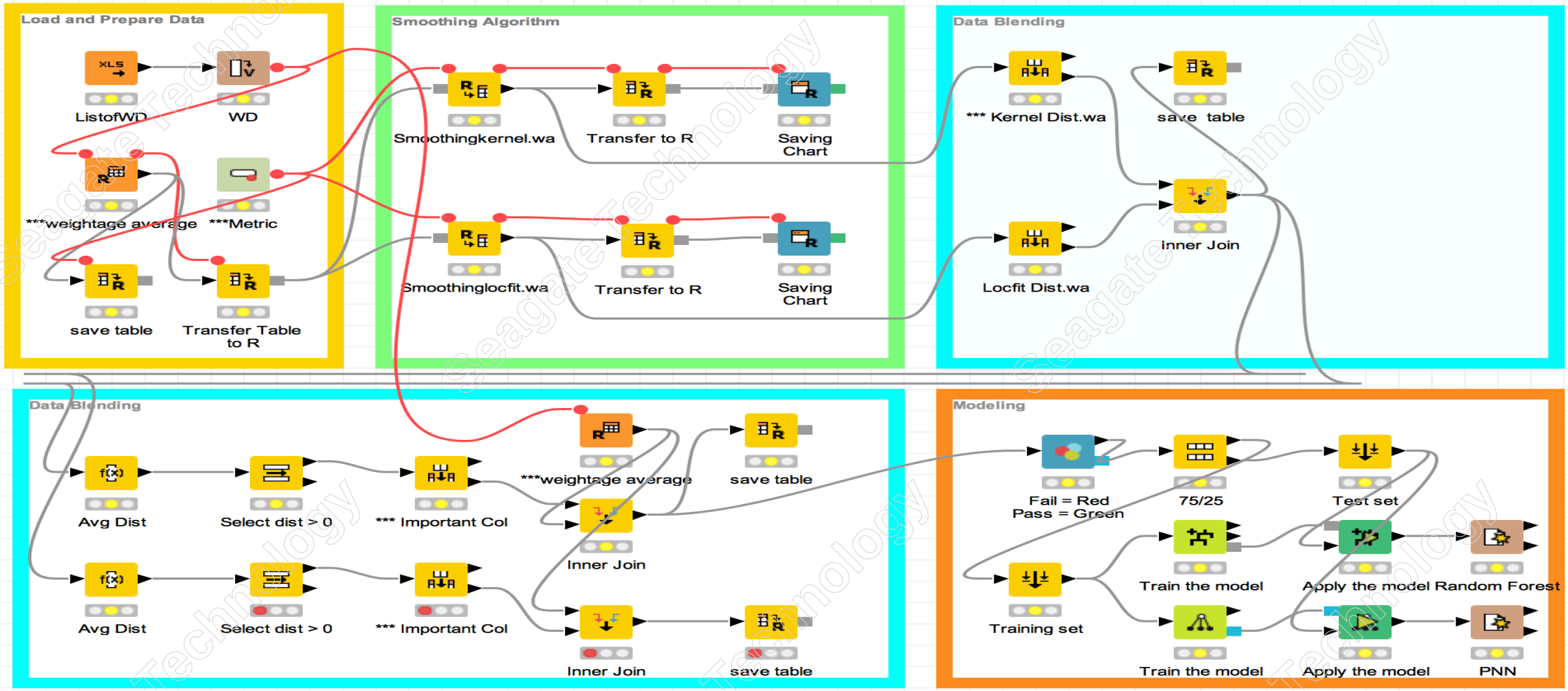
© Seagate Technology, LLC, 2018

KNIME Deployment Example 1 – Component Manufacturing Quality Assessment



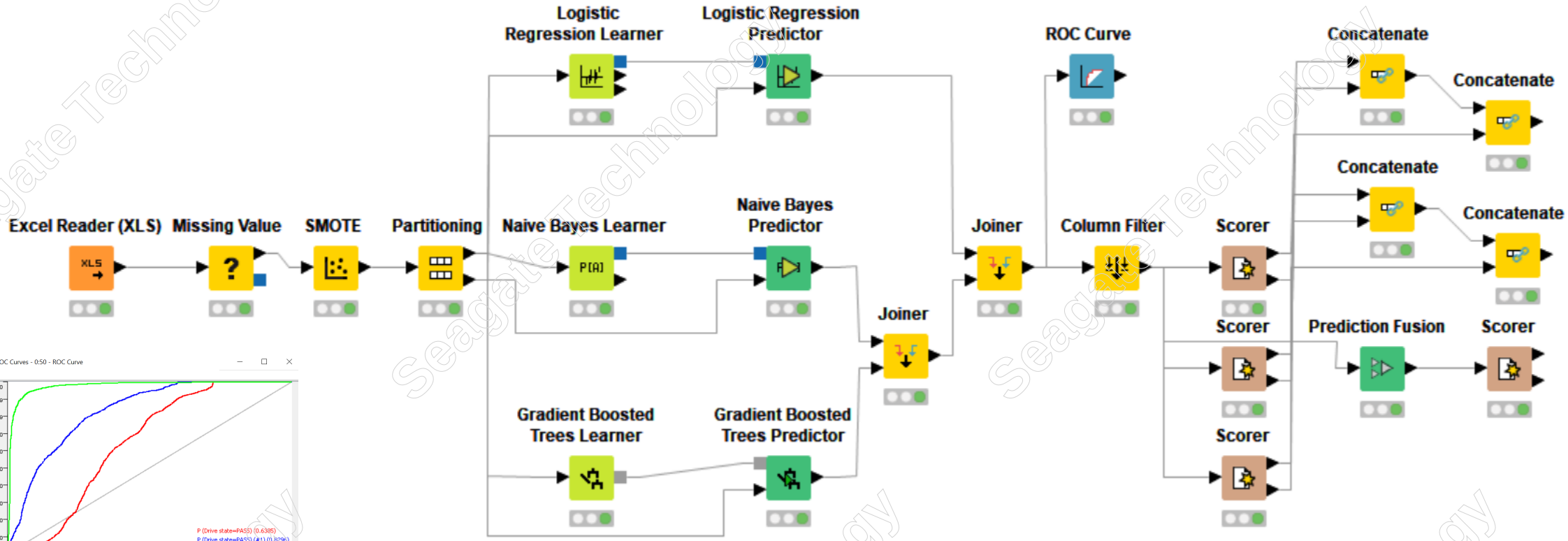
© Seagate Technology, LLC, 2018

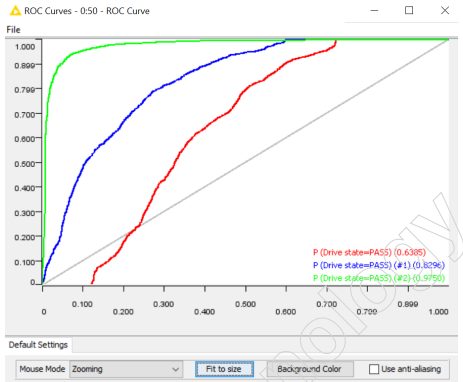
KNIME Deployment Example 1 – Component Manufacturing Quality Assessment



© Seagate Technology, LLC, 2018

KNIME Deployment Example 2 – Manufacturing Equipment Health Prediction

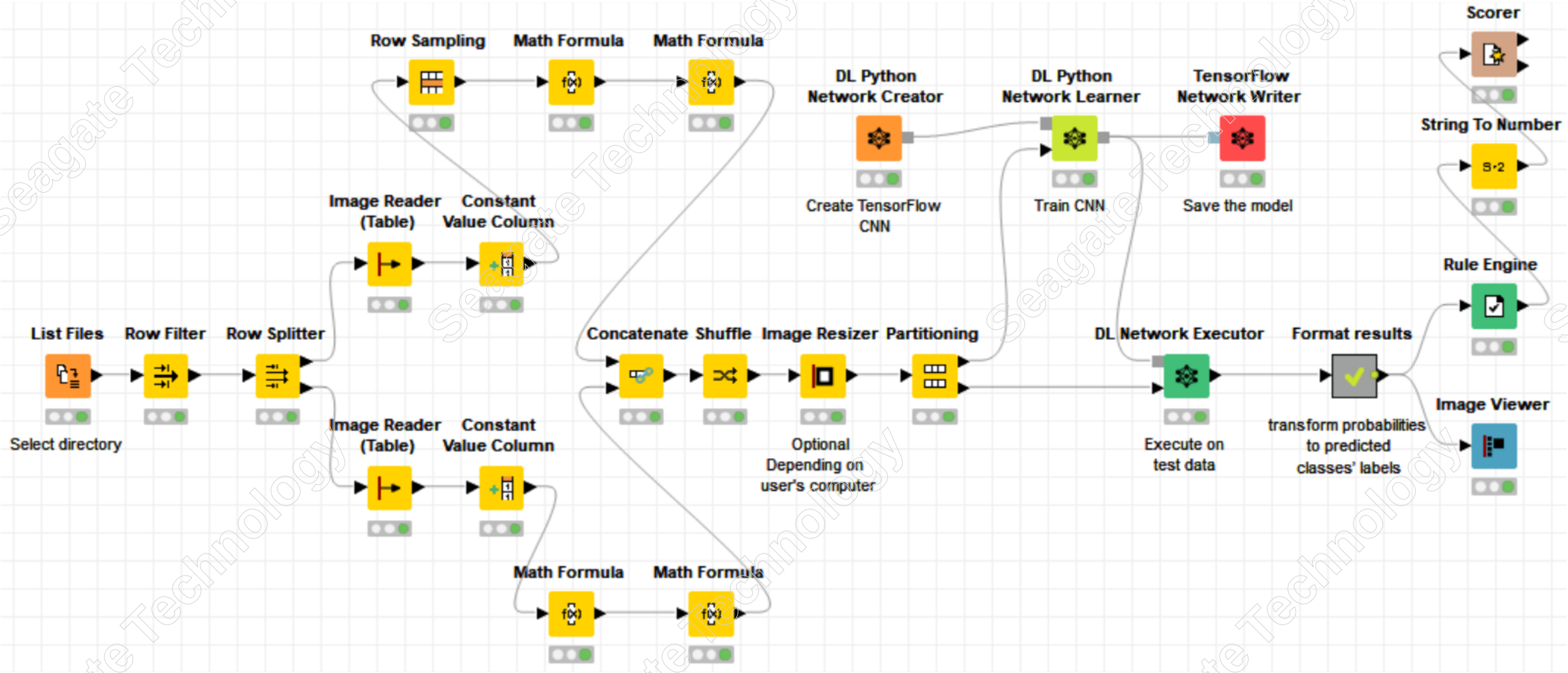




Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specifity	D F-meas...	D Accuracy	D Cohen'...
FAIL	1363	464	2356	1473	0.481	0.746	0.481	0.835	0.585	?	?
PASS	2356	1473	1363	464	0.835	0.615	0.835	0.481	0.709	?	?
Overall	?	?	?	?	?	?	?	?	?	0.658	0.316
FAIL_dup	1707	331	2489	1129	0.602	0.838	0.602	0.883	0.7	?	?
PASS_dup	2489	1129	1707	331	0.883	0.688	0.883	0.602	0.773	?	?
Overall_dup	?	?	?	?	?	?	?	?	?	0.742	0.484
FAIL_dup_dup	2792	128	2692	44	0.984	0.956	0.984	0.955	0.97	?	?
PASS_dup_dup	2692	44	2792	128	0.955	0.984	0.955	0.984	0.969	?	?
Overall_dup_...	?	?	?	?	?	?	?	?	?	0.97	0.939

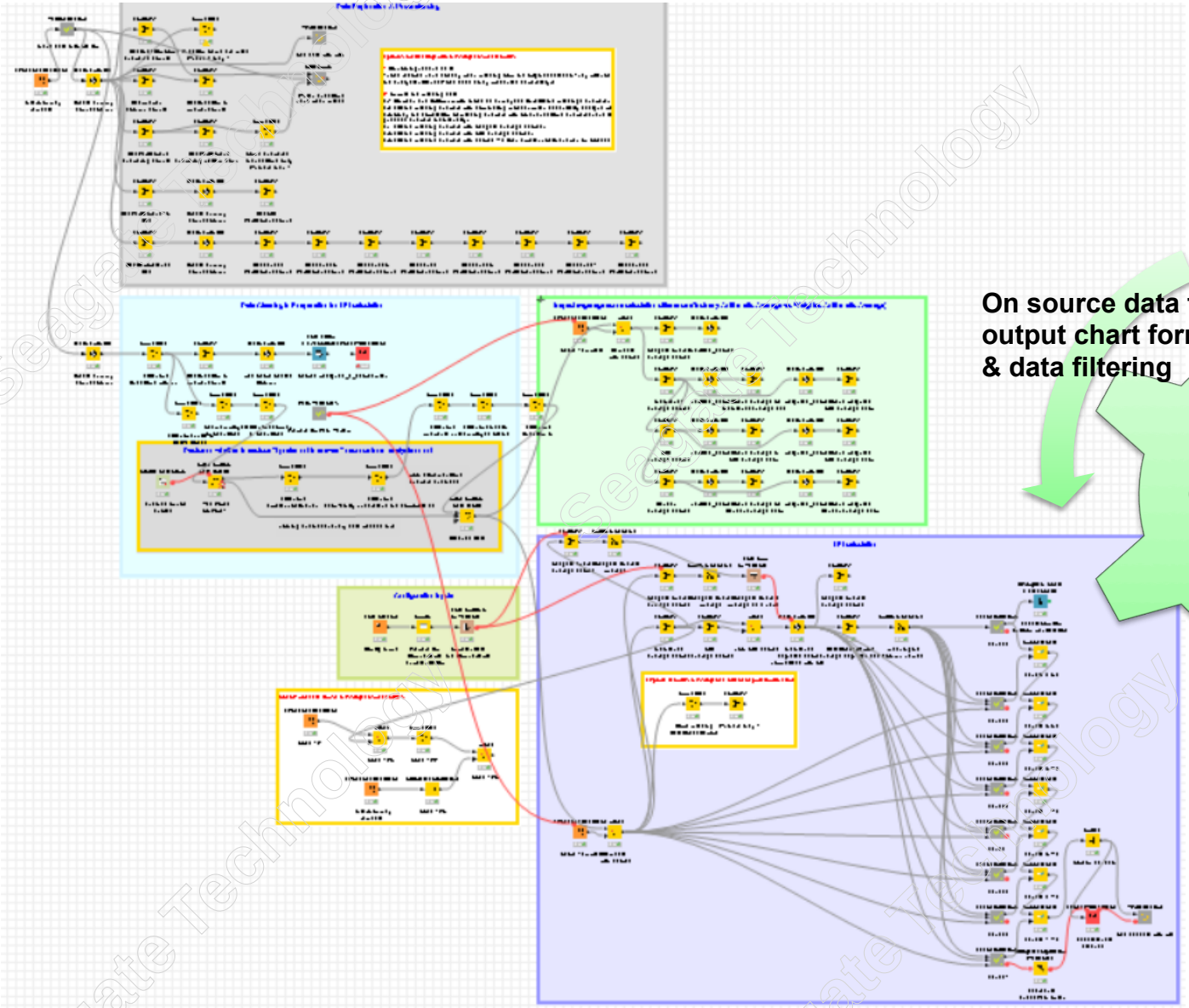
© Seagate Technology, LLC, 2018

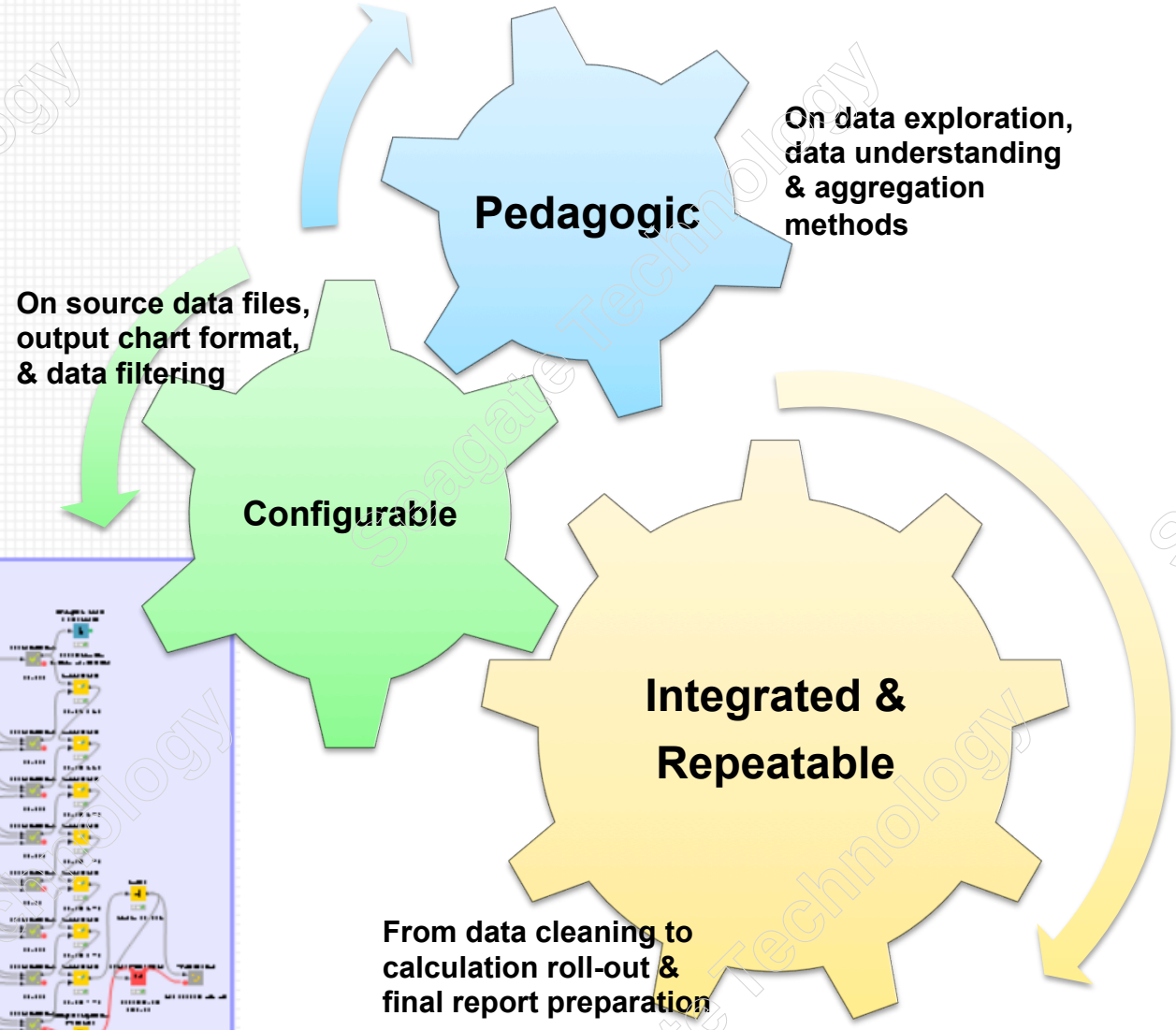
KNIME Deployment Example 3 – Manufacturing Image Analytics via Tensorflow Integration



© Seagate Technology, LLC, 2018

KNIME Deployment Example 4 – HR Analytics





© Seagate Technology, LLC, 2018

Summary

- **Guided Analytics Software implementation at Seagate**
  - Significant progress made over the past year
  - Many activities underway
    - KNIME software development and solutions deployment
    - Implementation areas: Manufacturing, Engineering, R&D, HR and other Business functions
  - KNIME server integration and rollout will further accelerate adoption

- **Application Areas**
  - 1) ETL, Data Query, 2) Integration Platform, 3) Modeling, Prediction

- **Benefits**
  - Enable our citizen data scientists and data analysts to do more with their data
  - Automation (e.g. reduce task duration: from days to minutes, seconds)

- **Our wish list to further enhance implementation**
  - Better visualization and dashboard features
  - More interactive data exploration
  - Interactive learning materials for newcomers

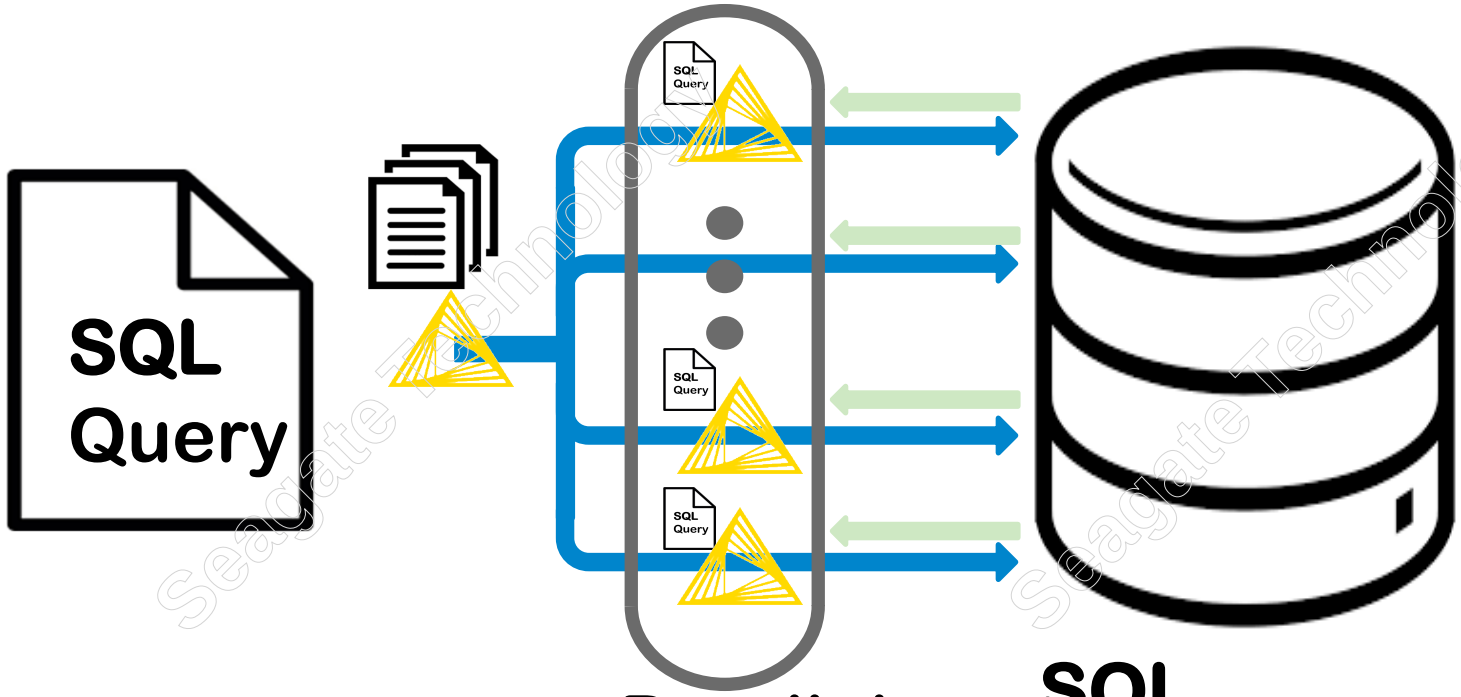
© Seagate Technology, LLC, 2018

KNIME Live Demo:
Parallel Execution of Data Sources in KNIME

Seagate Technology

© Seagate Technology, LLC, 2018

KNIME Live Demo: Parallel Execution of Data Sources in KNIME



Parallel SQL Execution

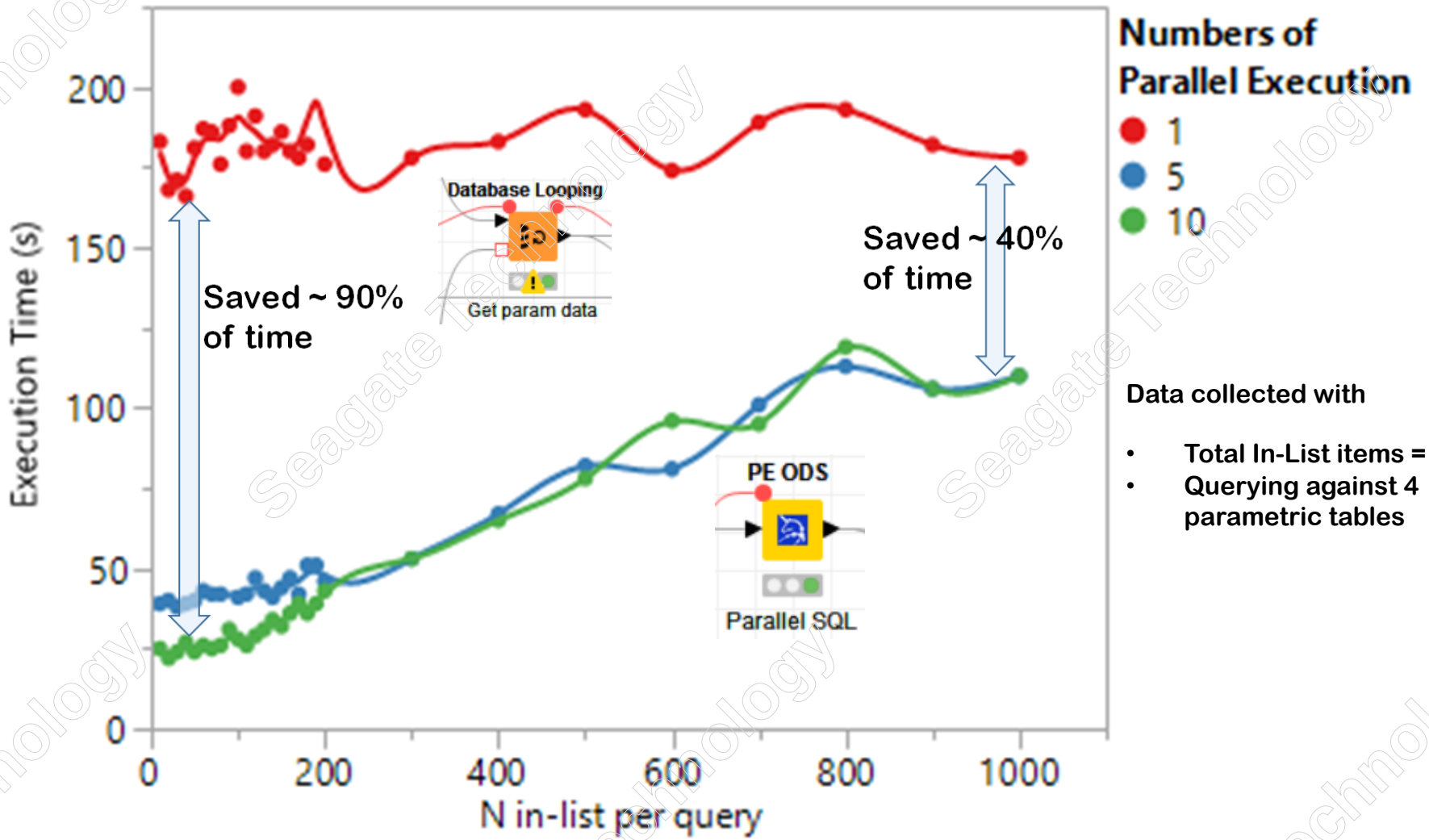
SQL Database

Multi-threading,
Not Multi-processing

Multitasking: Parallel Execution on Data Sources with KNIME/Java

© Seagate Technology, LLC, 2018

KNIME Live Demo: Parallel Execution of Data Sources in KNIME



Data collected with

- Total In-List items = 2000
- Querying against 4 parametric tables

The Power of Multithreading: Parallel Execution on Data Sources

© Seagate Technology, LLC, 2018