Classify Your Sensitive Data Assets Automatically Using IBM Watson Knowledge Catalog

LAB Session **5479**

—

Mallika Razdan
Technical Specialist
Data and AI | Public & Healthcare

Vishwanath Kamat
Lead Cloud Pak for Data Architect
Data and AI Expert Labs SWAT

Please note

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM’s sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

Notices and disclaimers

U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided. The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."

Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

Notices and disclaimers continued

It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

Contents

Lab overview
- Data Governance Challenges
- Overview of Data assets in question
- Potential ways to mitigate
- Watson Knowledge Catalog approach

Technology overview
- Platform architecture
- Key functionality used in the lab

Lab Setup

- Skytap environment access
- Lab guide and user IDs to use

Lab Overview

Data Governance Challenge

ABC Healthcare Inc is a , healthcare service provider across the United States. The Chief Data Officer(CDO) has been tasked to deal with patient data privacy compliance. The company has recently acquired another provider. Each organization has numerous data sources that need to managed and governed in compliance with federal as well as state regulations.

One of first task for CDO's office is to identify all PII (personally identifiable information) data assets across companies and data sources efficiently with limited human resources available.

Data assets under consideration

There are numerous sources that CDO's office is responsible for. As pilot for this project, following tables will be used from a db2 database

Patients
ID CHAR(40),
BIRTHDATE DATE,
DEATHDATE DATE,
SSN CHAR(11),
DRIVERS CHAR(20),
PASSPORT CHAR(20),
PREFIX CHAR(10),
FIRST CHAR(20),
LAST CHAR(20),
SUFFIX CHAR(20),
MAIDEN CHAR(20),
MARITAL CHAR(1),
RACE CHAR(20),
ETHNICITY CHAR(20),
GENDER CHAR(1),
BIRTHPLACE VARCHAR(80),
ADDRESS VARCHAR(150),
CITY CHAR(30),
STATE CHAR(40),
ZIP CHAR(10)

Allergies
START DATE,
STOP DATE,
PATIENT CHAR(40),
ENCOUNTER CHAR(40),
CODE CHAR(20),
DESCRIPTION CHAR(40)

Potential ways to mitigate

Capture all PII in data model
Data modeler can create and document PII info within data modeling tool.
Problem : Lot of manual effort and difficult to share across processes and workflows.

Capture all PII assets in spreadsheets
Ask all data owners to create their own ways to capture PII assets.
Problem : Lot of manual and error prone

Metadata tool
Capture all metadata across lines of businesses within a single metadata repository. Have LOBs owners to setup and configure PII with specific tags.
Problem: Varies by tool. Many tools will require extensive manual effort to classify data assets.

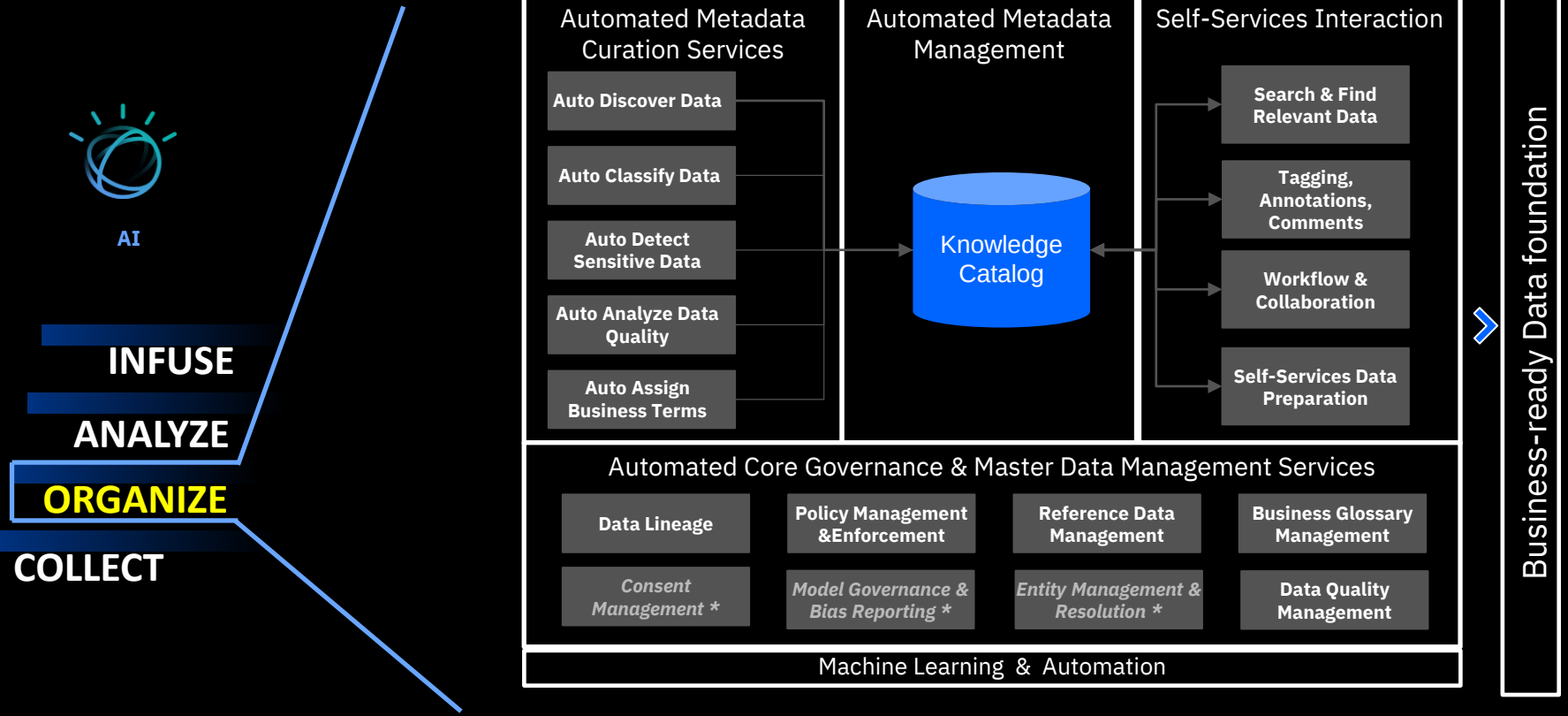
Do nothing !!

Watson Knowledge Catalog (WKC) approach

WKC provides single enterprise wide metadata tool that individual LOBs can collaboratively work with. The tool provides capturing of metadata automatically using machine learning(ML) based identification of data classes, assigning business glossary to technical metadata and governance workflows. The tool provides automation rules that significantly reduces human intervention in identification and management of PII data assets.

Watson Knowledge Catalog

End-To-End Fully Integrated Platform for Data Integration, Quality, Governance, And Consumption

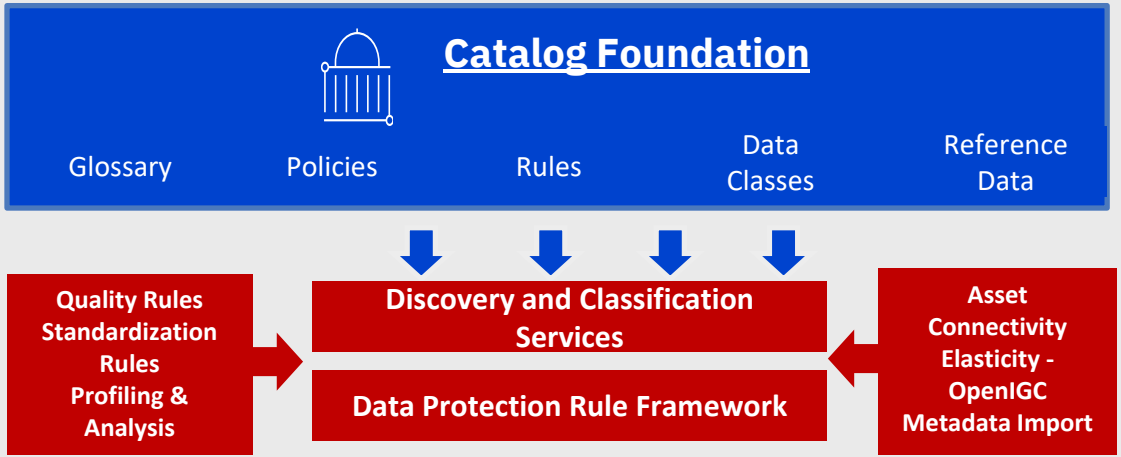


Watson Knowledge Catalog - Automated Governance

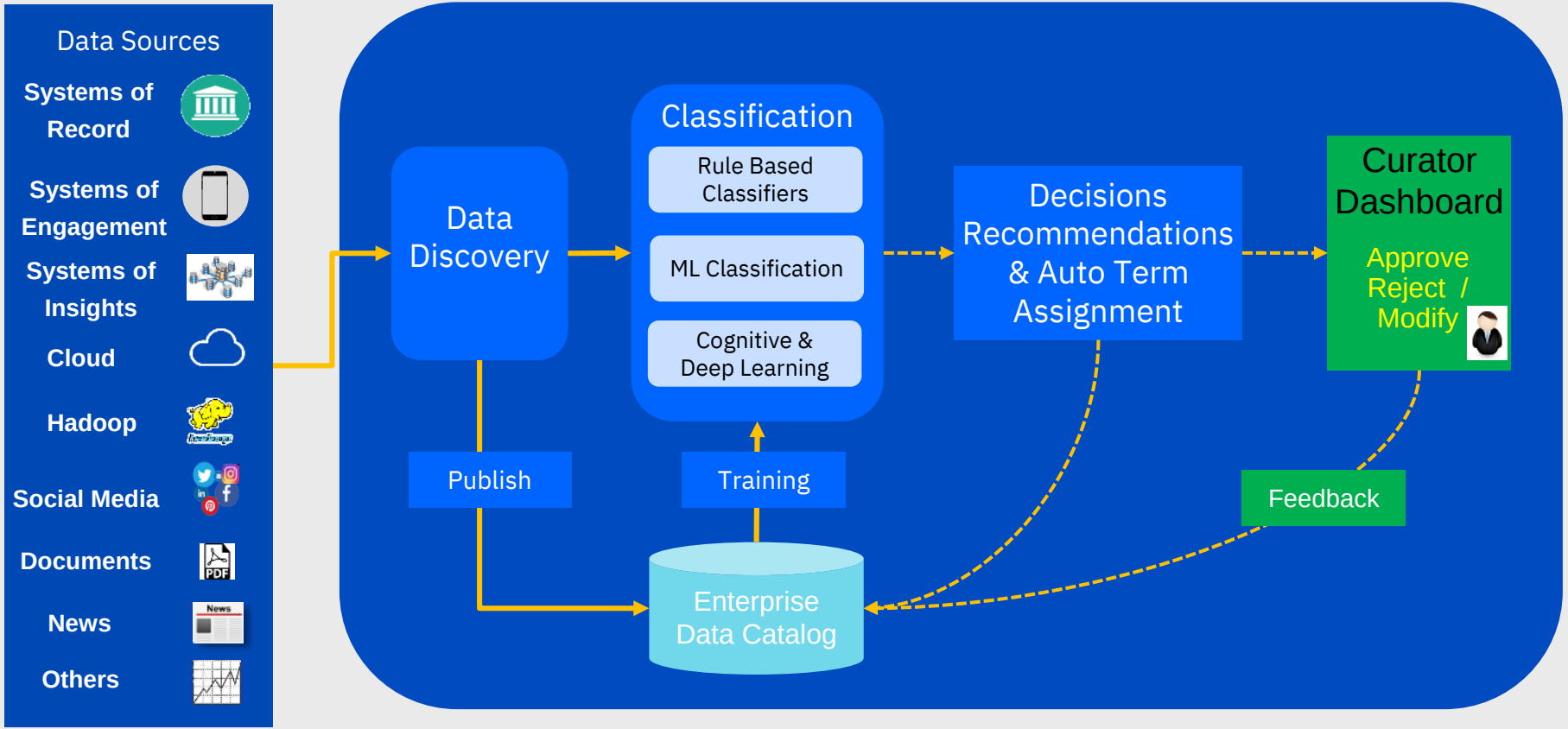
1.Build Catalog foundation through
- Business Glossary
- Governance Policies
- Protection Rules
- Data Classes
- Reference Data

2. Automated Discovery
- Setup Data Source Connection
- Initiate Import and Discovery Job
- Review and Publish Data Asset to catalog

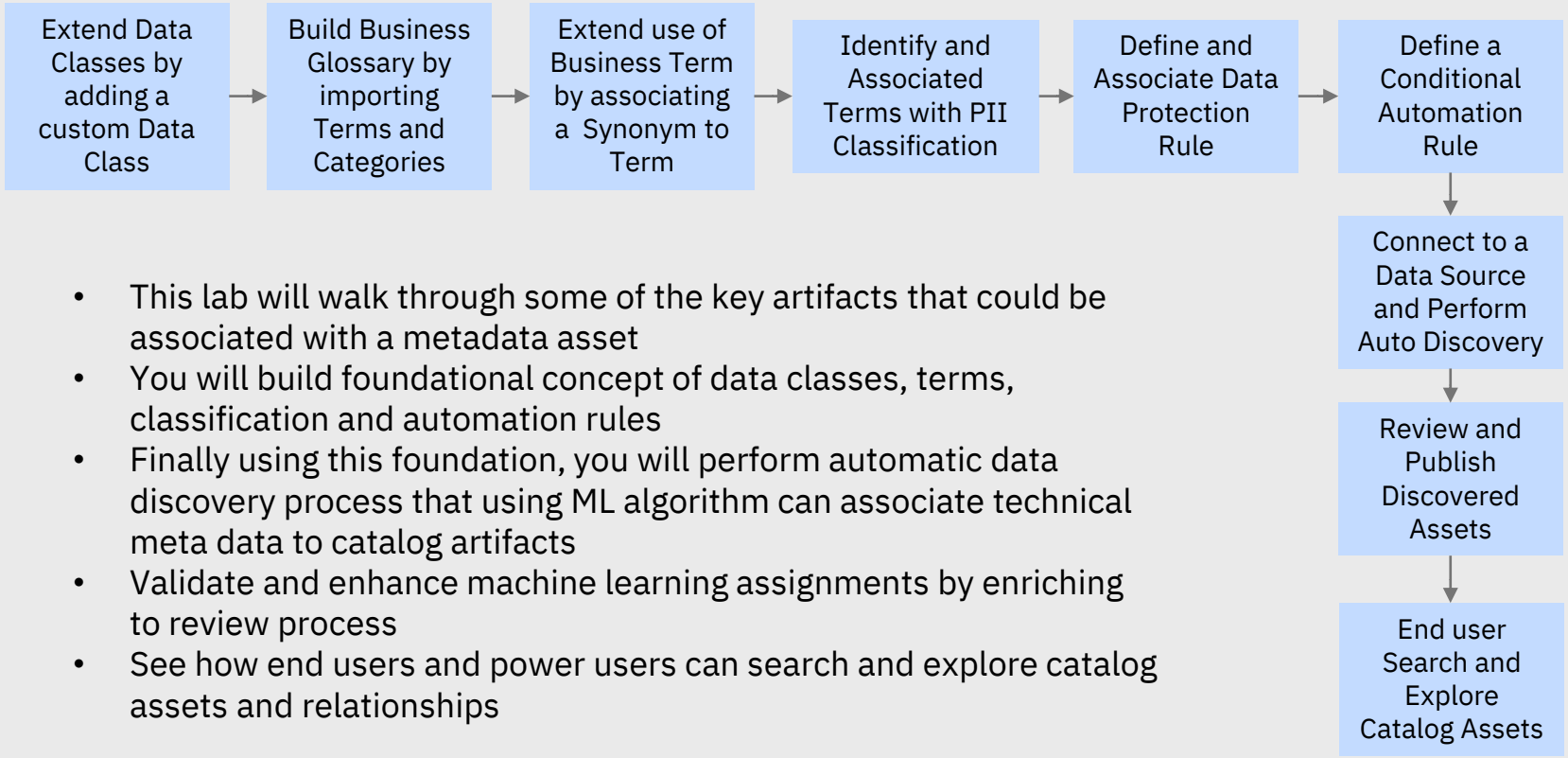
3. Consume Catalog Assets
- Enterprise Search
- Explore Relationship Graph



Auto Classification and Term Assignment



Lab Flow



- This lab will walk through some of the key artifacts that could be associated with a metadata asset
- You will build foundational concept of data classes, terms, classification and automation rules
- Finally using this foundation, you will perform automatic data discovery process that using ML algorithm can associate technical meta data to catalog artifacts
- Validate and enhance machine learning assignments by enriching to review process
- See how end users and power users can search and explore catalog assets and relationships

Lab Setup

Lab document and presentation deck is available at this public git location

https://github.com/vishkamat/think2020_lab5749

- WKC is installed as part of Cloud Pak for Data instance. Skytap is a lab hosting service being used.
- All end user activity will be through Windows virtual machine through browser.
- There are 4 instances of Cloud Pak for data, each with user1 through user10 pre-created
- You will be assigned an instance to login with your user ID and password

- Each instance includes a db2WH database instance which has Patient and Allergies tables created
- A database connection has been created with DB2THINK2020 name
- Since concurrent Auto Discovery jobs may be queued due to concurrent access within same instance, a job has been run and available for you to run through subsequent steps to complete this lab.

Thank you

Mallika Razdan
Technical Specialist
Data and AI | Public & Healthcare
mrazdan@ibm.com

Vishwanath Kamat
Lead Cloud Pak for Data Architect
Data and AI Expert Labs SWAT
vkamat@us.ibm.com


IBM