Module 3

Discovery

Datastores

ما هو الـ Data Store؟

هو عبارة عن كيان يُمثل **اتصال بمصدر بيانات**. يستخدمه DPM لتنفيذ عمليات مثل:

- المسح (Scan)
- التحليل (Profiling)
- تصنيف البيانات الحساسة (Classification)

أنواع مصادر البيانات المدعومة:

- تطبيقات السحابة (Cloud Applications)
  - Cloudera Navigator
- قواعد البيانات العلائقية (Relational Databases)
- أنظمة الملفات (File Systems)
  - Hadoop
- مستودع PowerCenter
- خدمات Data Engineering Integration من إنفورماتيكا
- خدمات تطبيقات السحابة من إنفورماتيكا
  - SQL Server Integration Services (SSIS)

خطوات إنشاء Data Store:

1. **أولاً يجب إنشاء**:
   - المواقع (Locations)
   - مجموعات مصادر البيانات (Data Store Groups)
2. **طرق إنشاء Data Store**:
   - يدويًا (Manually)
   - استيراد من PowerCenter Repository

- استيراد من ملف CSV

ما الذي يتم تحديده عند الإنشاء؟

- معلومات الاتصال (Host, Port, Username...)
- ربط الـ Data Store بـ:
  - الموقع (Location)
  - مجموعة Data Store
  - مجموعة الأمان (اختياري)

بعد الإنشاء:

- يمكن إضافة Data Store إلى **وظائف الفحص (Scan Jobs)**
- يقوم DPM بالاتصال بالمصدر واكتشاف **البيانات الحساسة** و**إنتاج التقارير**

أداة الاستيراد والتصدير (Import/Export Utility) في DPM

أداة الاستيراد (Import Utility)

تُستخدم لاستيراد البيانات التالية عبر ملفات CSV:

- تفاصيل اتصال الـ **Data Stores**
- **أسماء مستعارة** (Aliases) للـ Data Stores
- **ملاك البيانات** (Data Owners)
- حالة الحماية (Protection Status)
- القوائم السوداء والبيضاء العالمية (**Global Blacklist & Whitelist**)
- أخرى Proliferation مخصصة من أدوات **Data Lineage**

أداة التصدير (Export Utility)

تُستخدم لتصدير البيانات التالية:

- تفاصيل اتصال الـ Data Stores

- ملاك البيانات (Data Owners)
- القوائم السوداء والبيضاء
- إعدادات قواعد البيانات (DB Configuration)

Import Lineage في DPM

ما هو Import Lineage؟

- إمكانية **استيراد معلومات الـ Data Lineage** من ملفات بصيغة **CSV**
- تُضاف المعلومات المستوردة إلى **مستودع DPM (Repository)**

الفكرة:

- أحيانًا تكون معلومات تدفق البيانات (lineage) موجودة في أدوات خارجية مثل:

- **Talend**
- **IBM DataStage**

- يمكنك **تصدير lineage من هذه الأدوات** إلى ملف CSV
- ثم **استيراده في DPM** لتعزيز تتبع البيانات

ملاحظات إضافية:

- يمكنك تحديد إذا كان العمود **محمي** (Masked أو Encrypted)
- بعد إجراء **Scan**، يقوم DPM بدمج نتائج الفحص مع الـ Lineage المستورد
- يتم عرض **انتشار البيانات الحساسة (Proliferation)** بين الـ Data Stores في الـ Dashboard

---

مثال:

Source	Target	Column	Protected
db1.tableA	db2.tableB	SSN	Yes

Import Sensitivity & Protection Status

ما الهدف من الاستيراد؟

- استيراد **حالة الحساسية (Sensitivity)** و**الحماية (Protection Status)** للأعمدة من ملف CSV.
- يُستخدم عندما تكون لديك **معرفة مسبقة** بالأعمدة الحساسة أو غير الحساسة.

ما الذي يمكن استيراده؟

- **اسم العمود**
- هل العمود:
- 🟢 **Whitelist** = دائمًا حساس (مثل SSN، رقم بطاقة)
- 🔴 **Blacklist** = غير حساس إطلاقًا، مهما كان محتواه

أين يُخزن هذا؟

- يتم تخزين الحالة في **DPM Repository**.
- تُستخدم أثناء عمليات الفحص (Scan) لتحديد الأعمدة تلقائيًا حسب الحساسية.

مثال CSV:

DataStoreName, TableName, ColumnName, SensitivityStatus
HR_DB, Employees, SSN, Whitelist
HR_DB, Employees, Notes, Blacklist`

الفائدة:

- تقلل الحاجة لفحص يدوي متكرر
- تزيد دقة تحديد الأعمدة الحساسة في نتائج الفحص
- تُستخدم لاحقًا في السياسات الأمنية والتقارير

Merge & Test Data Stores

Merge Data Stores

- الهدف: دمج **Data Stores مكررة** تتصل بنفس مصدر البيانات.
- ✅ يساعد على:
  - منع تكرار الفحص (Scan) لنفس المصدر.
  - إعطاء **عدد دقيق** لمصادر البيانات في الـ Dashboard.
- يتم الدمج بعد الاستيراد أو الإنشاء اليدوي لمصادر بيانات متكررة.

Test Connection

- يمكنك اختبار الاتصال بمصدر البيانات عند:
  - إنشاء Data Store.
  - أو تعديل الإعدادات الخاصة به.
- مدعوم لكل أنواع Data Stores **ما عدا ODBC**.

Scans in DPM

ما هو الـ Scan؟

- **كائن (Object)** داخل DPM Repository.
- يُستخدم لـ **اكتشاف وتصنيف البيانات الحساسة** في مصادر البيانات (Data Stores).
- عند تشغيله، يُنشئ **Scan Job** لكل Data Store مشارك.

مكوّنات Scan

- تعريف لمجموعة من إعدادات المسح.
- يتضمن:

- Data Stores
- Data Domains
- Classification Policies

- جدول زمني أو تنفيذ فوري

وقت الاستخدام:

- من خلال **Scans Workspace**:
  - إنشاء Scan جديد.
  - تشغيل Scan فورًا أو جدولته.
  - إدارة نتائج الفحص السابقة.

خيارات التنفيذ:

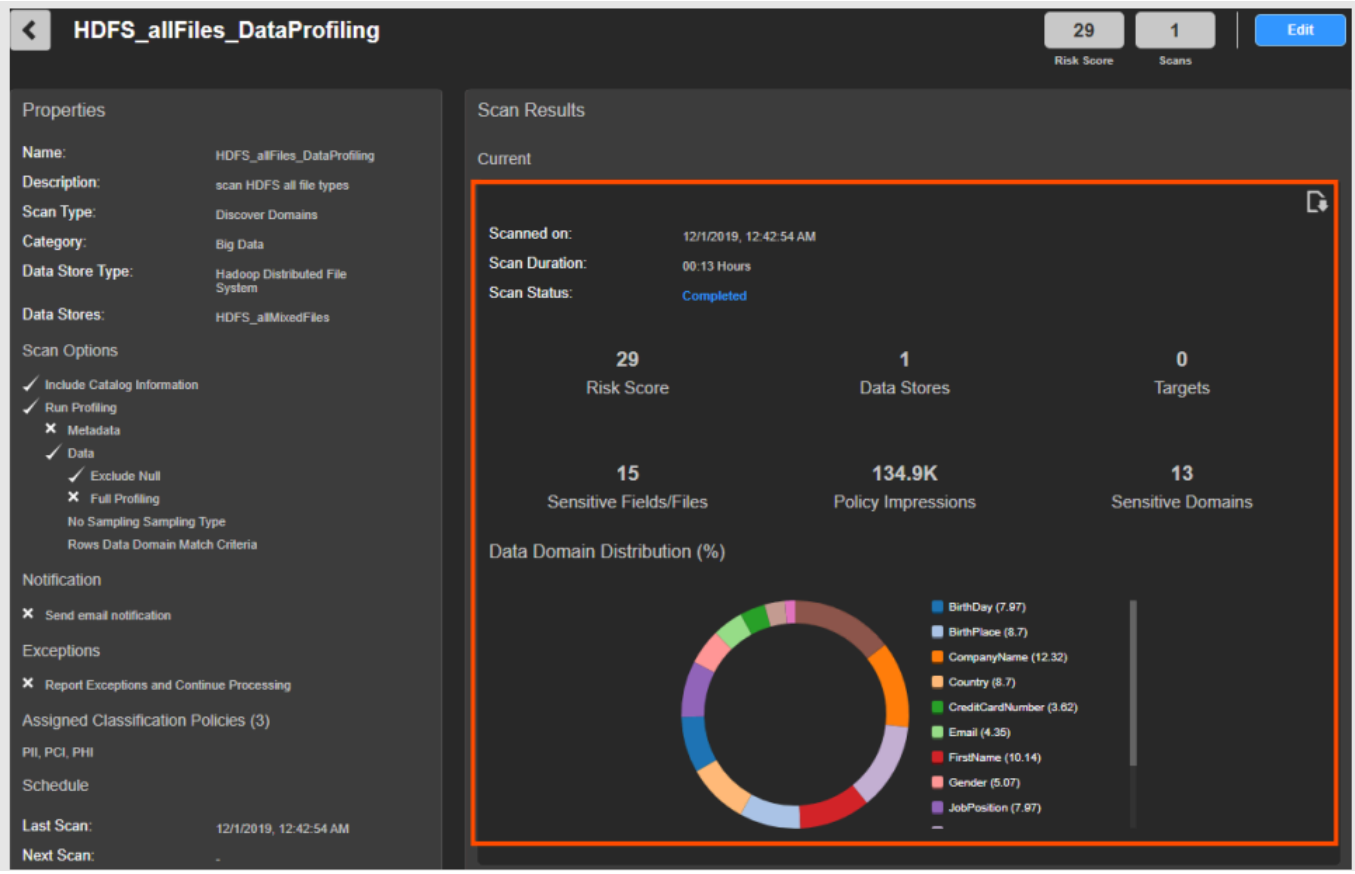
- **Run Now**: تنفيذ فوري.
- **Schedule**: جدولة الفحص ليُنفّذ لاحقًا.

مثال عملي:

Scan Name: HR_Scan
Data Stores: HR_DB, Payroll_DB
Policies: PII, PHI
⇨ DPM ينشئ:

- Job لـ HR_DB
- Job لـ Payroll_DB`

Scans Details Page



متى تقدر تحذف Scan؟

تقدر تحذف الـ Scan **لو حالته كانت واحدة من الحالات التالية**:

- ✅ **Completed** (اكتمل)
- ✅ **Scheduled** (مجدول)
- ✅ **Terminated** (تم إنهاؤه يدويًا)

❌ متى لا يمكنك الحذف؟

لا يمكن حذف الـ Scan إذا كانت حالته:

- ❌ **Failed** (فشل)
- ❌ **Paused** (مؤقتًا موقوف)
- ❌ **Stopped** (تم إيقافه أثناء التنفيذ)

Scan Management – خطوات إنشاء Scan

###$ الهدف:

تكوين **كيف ومتى** يتم فحص البيانات (Scan) لتحديد وتصنيف البيانات الحساسة.

Step 1: Define and Connect

- تختار **نوع الـ Scan** (مثلاً Database Scan).
- تحدد **الـ Data Stores** اللي عايز تفحصها.
- لازم تكون الـ Data Stores مربوطة ومتعرفة مسبقًا.

Step 2: Assign Classification Policy

- تعيّن **سياسات التصنيف** (Classification Policies) للـ Scan.
- السياسات دي تحتوي على مجموعة من **Data Domains** (مثلاً: رقم بطاقة، رقم قومي).
- كل Policy ممكن يكون فيها **مستوى حساسية** و**تكلفة المخاطرة**.

> **Note**
>
> PII & PCI & PHI

Step 3: Schedule

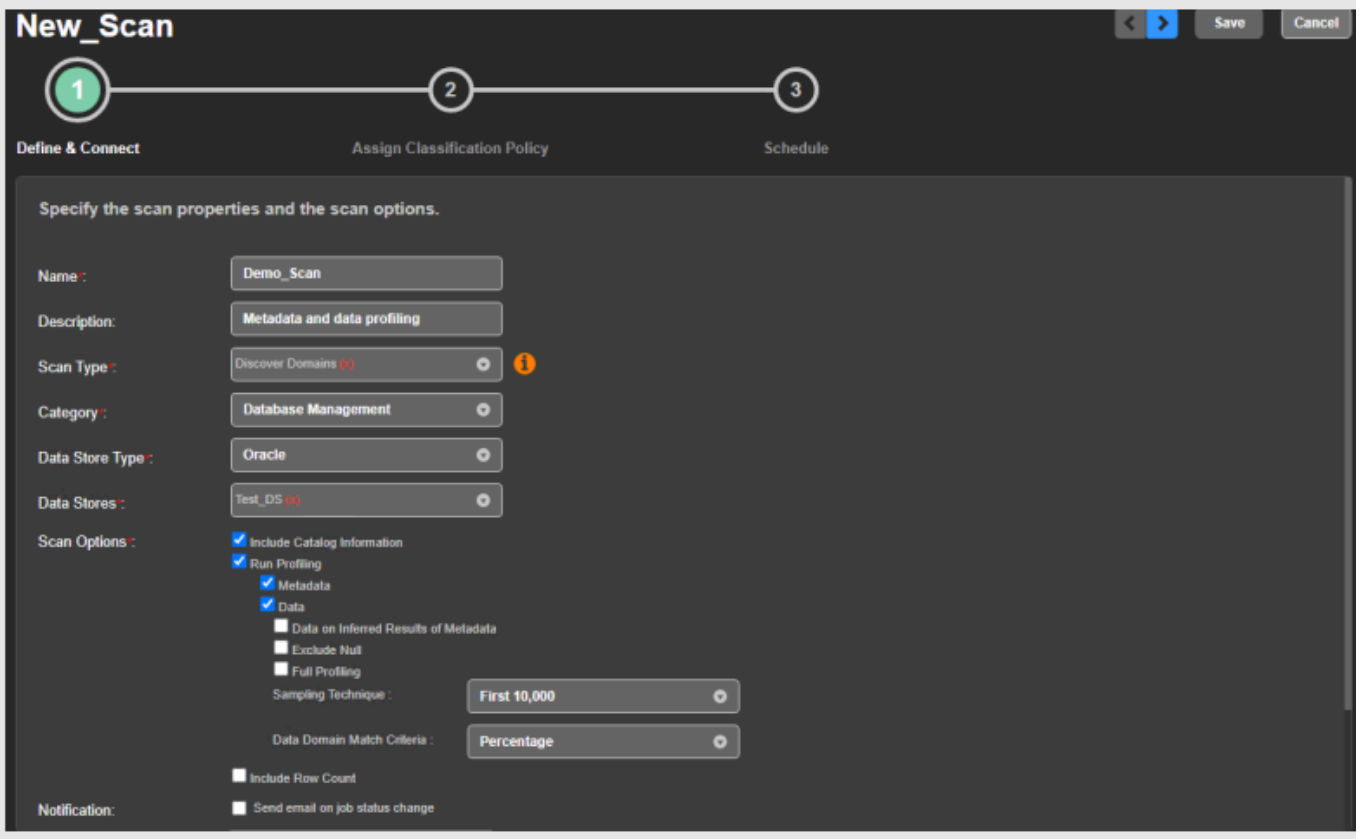
- تحدد **متى** عايز الـ Scan يتنفذ:
  - فورًا (Immediately)
  - في وقت محدد (Scheduled)
  - بشكل متكرر (Recurring)

مثال عملي:

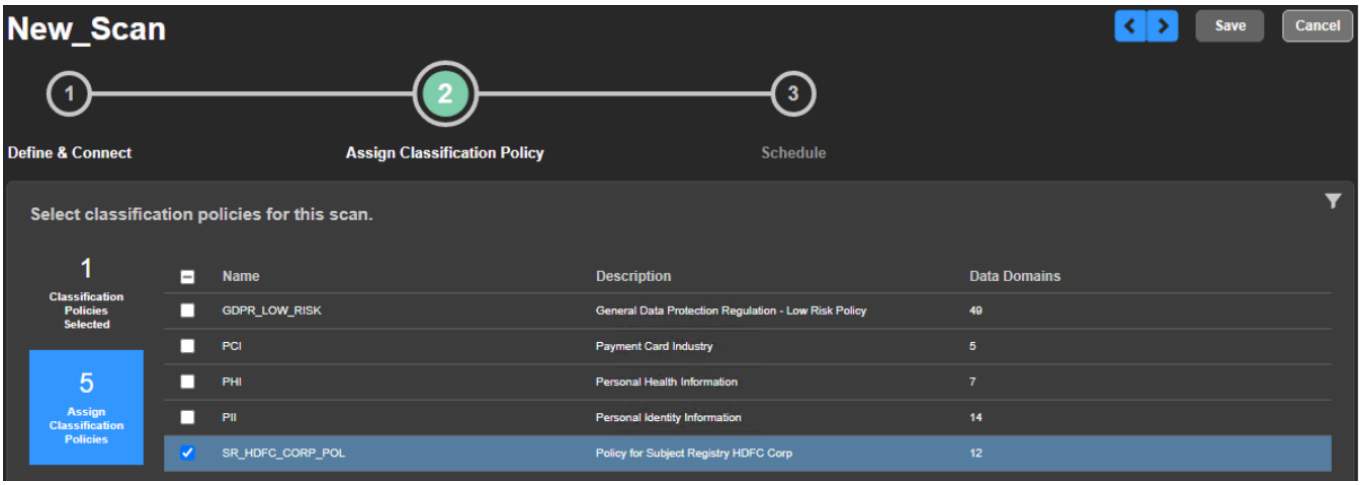
Scan Name	Data Store	Policy	Schedule
HR_Scan	HR_DB	PII Policy	Every Monday 10 AM

Scan Name	Data Store	Policy	Schedule
Sales_Scan	Sales_DB	PCI Policy	Run Now

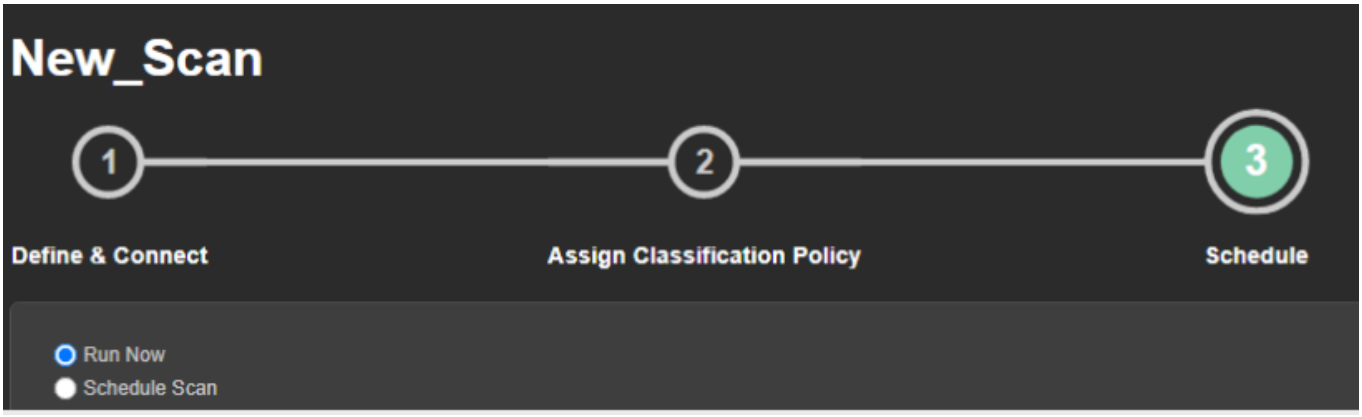
Step 1 - Define and Connect



Step 2 - Assign Policy



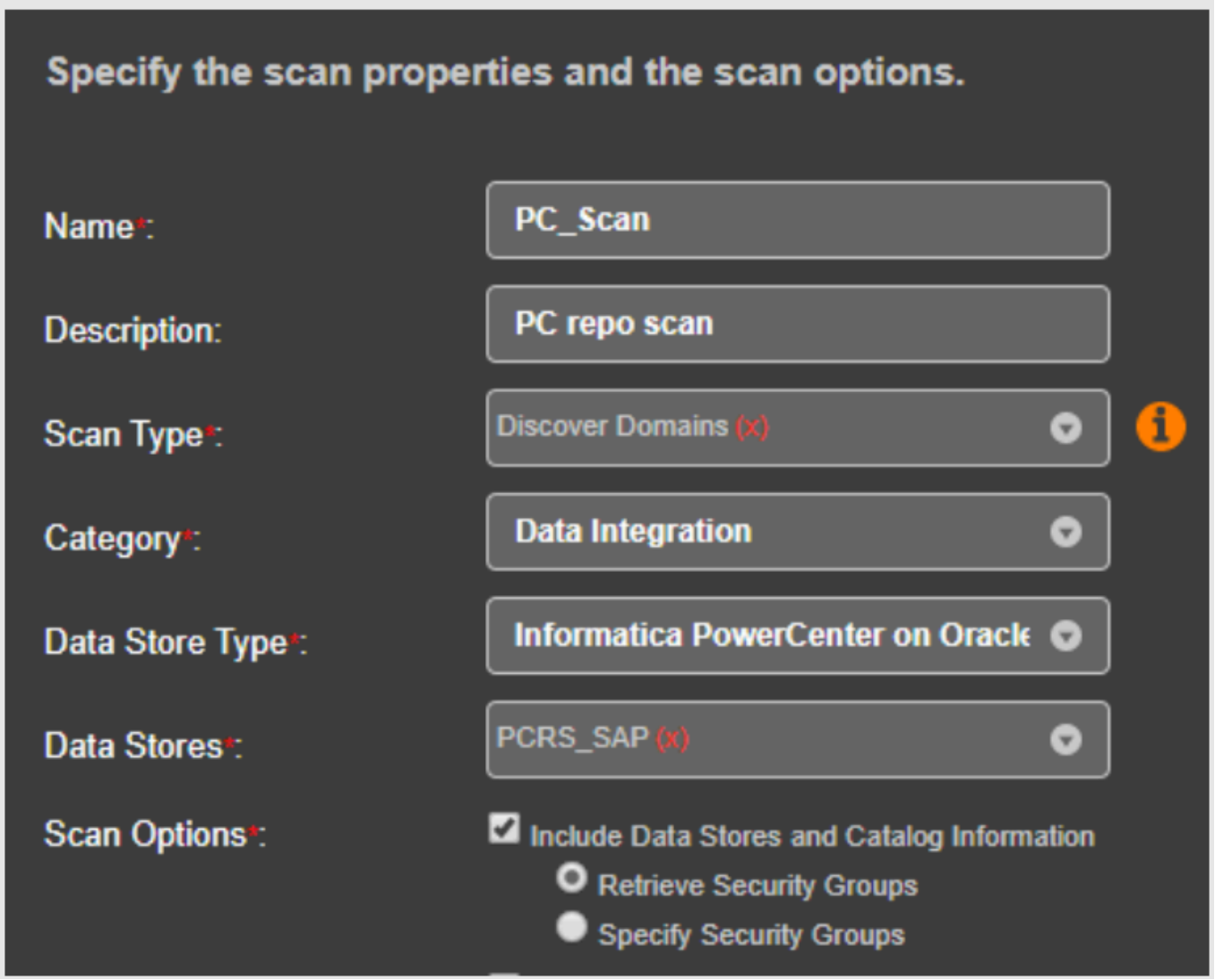
Step 3 - Schedule



Data Store Types

- PowerCenter Repository Service
- Relational Database
- Informatica Big Data Management
- Informatica Cloud Service
- Cloud Applications (Salesforce)
- Cloudera Navigator
- Hadoop
- File System
- SQL Integration Services

1- PowerCenter Repository Service



1. تكتشف كل مصادر البيانات (Data Stores) اللي PowerCenter بيشتغل عليها.
2. تحدد فين البيانات الحساسة (PII زي الرقم القومي، الإيميل، رقم الحساب).
3. تعرف البيانات دي **انتشرت منين لفين** — يعني سريانها داخل الـ Workflows (ده اسمه **Proliferation**).

خطوات تشغيل PowerCenter Repository Scan

الخطوة 1: شغّل Scan مع اختيار:

```
Include Data Stores and Catalog Information
```

اللي بيحصل:

- DPM يتصل بـ PowerCenter Repository.

- يقرأ كل الـ **Workflows, Sessions, Sources, Targets...** إلخ.
- يكتشف الـ **Data Stores** اللي PowerCenter بيتعامل معاها.
- يجمع الـ **metadata** بتاعتها (يعني أسماء الجداول، الأعمدة، البيانات المتداولة...).

🎯 **نتيجة الخطوة دي:**
أنت دلوقتي عندك في DPM كل مصادر البيانات (Data Stores) اللي PowerCenter بيشاور عليها.

الخطوة 2: شغّل Database Scan على كل Data Store

💡 ليه؟ لأن الخطوة الأولى جابتلك metadata بس، لكن مفيش تحليل حقيقي للبيانات نفسها.

اللي بيحصل:

- DPM يدخل فعليًا على قواعد البيانات نفسها.
- يقرأ القيم اللي جوه الأعمدة.
- ويطبق عليها السياسات (Policies) والـ Data Domains (زي: رقم قومي، إيميل...).

🎯 **نتيجة الخطوة دي:**
يعرف إذا كان العمود ده يحتوي على بيانات حساسة ولا لأ (مثلًا 90% من العمود فيه أرقام قومية ← يعتبره PII).

الخطوة 3: شغّل Scan تاني مع اختيار:

```
Identify Proliferation and Data Protection
```

اللي بيحصل:

- DPM 2 ياخد البيانات الحساسة اللي اكتشفها في الخطوة.
- ويبدأ يتتبعها داخل الـ PowerCenter workflows:
  - راحت منين لفين؟
  - هل وصلت لـ Target جديد؟
  - هل فيه Transformation حصل؟

🎯 **نتيجة الخطوة دي:**
تحصل على **خريطة انتشار البيانات الحساسة** داخل البنية اللي بيشتغل عليها PowerCenter.

مثال واقعي للتوضيح:

تخيل Workflow اسمه: `Employee_Data_ETL`

- Source: Oracle DB → فيه جدول فيه عمود اسمه `National_ID`
- Data Warehouse بينقل العمود ده إلى Workflow

اللي بيحصل:

1. DPM يشوف الـ Workflow ويلاقي إن فيه نقل لعمود `National_ID`.
2. يعمل Scan ويشوف إن القيم فعلاً كلها أرقام قومية → يعتبره **PII**.
3. يشوف إن العمود ده اتنقل لـ Data Warehouse → **Proliferation حصل**.
4. يسجله في **خريطة الانتشار (Proliferation Map)**.

ملخص الخطوات:

الخطوة	الوظيفة	النتيجة
1	Scan بـ "Include Data Stores"	استيراد metadata و data stores من PowerCenter
2	Database Scan	تحليل القيم وتحديد الأعمدة الحساسة
3	Scan بـ "Identify Proliferation"	رسم خريطة انتشار البيانات الحساسة

ليه ده مهم؟

- لأنك بتقدر تعرف فين بياناتك الحساسة راحت، وده يساعدك:
  - تأمّنها.
  - تطبّق سياسة حماية مناسبة.
  - تمنع تسريب أو سوء استخدام.

2- Database Scans

إيه هو Database Scan؟

هو نوع من أنواع الـ Scans في DPM، بيستخدم عشان:

1. يكتشف الأعمدة اللي ممكن تكون فيها **بيانات حساسة** (PII، PCI...).
2. يحسب عدد الصفوف اللي فيها بيانات حساسة.

3. يستخرج المعلومات (Metadata) من قاعدة البيانات زي أسماء الجداول والأعمدة.

خطوات تنفيذ Database Scan (مكوناته الرئيسية):

1. Extract Metadata

- DPM يوصل لقاعدة البيانات.
- يجيب معلومات عن:
  - أسماء الجداول
  - أسماء الأعمدة
  - أنواع البيانات (Data Types)

> ده بيساعد إنه يفهم هيكل قاعدة البيانات.

---

2. Perform Data Domain Discovery (Profiling)

- هنا DPM يبدأ يحلل **بيانات الأعمدة نفسها** (القيم جوه الأعمدة).
- يستخدم الـ **Data Domains** زي:
  - National ID
  - Credit Card
  - Email

> ويبدأ يطابق القيم باستخدام:

- Regular Expressions
- Reference Tables
- Rules

> علشان يعرف إذا العمود ده فيه بيانات حساسة ولا لأ.

---

3. Calculate Row Count of Sensitive Columns

- بعد ما يلاقي عمود فيه بيانات حساسة، بيحسب:
  - **عدد الصفوف** اللي فيها بيانات فعلاً حساسة (مطابقة للـ Domain).

> 🎯 وده مهم لتحديد مدى خطورة العمود (Risk Score)

إعدادات مهمة:

Profiling Options:

- ممكن تختار إن الـ Scan يشتغل على:
- **Metadata** (اسم العمود فقط) **بس**.
- **Data** (القيم فقط) **بس**.
- **الاتنين معًا** (وده الأدق والأفضل غالبًا).

هل لازم أشغل كل حاجة مرة واحدة؟

مش شرط!

- ممكن تعمل Scan أول مرة بـ:
- **Include Catalog Information** → يجيب metadata فقط.
- ثم تعمل Scan تاني بـ:
- **Profiling** → يحلل القيم ويكتشف البيانات الحساسة.
- ثم تعمل Scan تالت بـ:
- **Include Row Count** → يحسب عدد الصفوف الحساسة.

أو ببساطة: تعمل Scan واحد بكل الخيارات مرة واحدة لو السيستم قوي.

الناتج النهائي من Database Scan:

المعلومة	معناها
Metadata	وصف الجداول والأعمدة
Sensitive Columns	الأعمدة اللي فيها بيانات حساسة
Sensitive Row Count	عدد الصفوف اللي فعلاً فيها بيانات حساسة
Data Domain Match	نوع البيانات الحساسة (رقم قومي؟ بطاقة؟...)

3- Informatica Big Data Management Scan

إيه هو BDE Scan؟

هو نوع من الـ Scans بيشتغل على **Informatica Big Data Edition (BDE)**، والهدف منه:

1. استيراد معلومات الـ Data Stores الموجودة في الـ Big Data Edition.
2. اكتشاف وتصنيف البيانات الحساسة.
3. تحليل انتشار البيانات الحساسة (Proliferation) داخل الـ **Mappings**.
4. التأكد من وجود حماية (Masking) على البيانات الحساسة باستخدام **DMO Transformations**.

خطوات عملية الـ Scan

الخطوة 1: **Run BDE Repository Scan — Include Data Stores**

- يروح لـ DPM ويحلل الـ Big Data Edition الخاصة به **metadata**.
- يستورد كل الـ **Data Stores** والمصادر اللي بيشتغل عليها.

🎯 *زي ما بتعمل مع PowerCenter Scan – هنا بنبني صورة عن بنية النظام.*

الخطوة 2: **Run Database Scan**

- بعد ما عرفنا الـ Data Stores من BDE، نبدأ نمسح البيانات الفعلية.
- الهدف هنا:
  - اكتشاف البيانات الحساسة باستخدام السياسات.
  - تحديد الأعمدة اللي فيها مثلًا أرقام قومية، إيميلات، بطاقات ائتمان...

📊 *نفس فكرة الـ Database Scan العادي.*

الخطوة 3: **Run BDE Repository Scan — Identify Proliferation and Protection**

- هنا DPM يحلل:
  - الـ **Mappings** جوه BDE: البيانات راحت فين؟ اتنقلت فين؟
  - هل البيانات الحساسة دي اتنقلت من جدول A لـ B؟
  - هل تم تطبيق **Data Masking** عليها باستخدام DMO Transformation؟

💡 *DMO (Dynamic Masking Object) ده نوع من التحويلات (Transformations) بيغلف البيانات أو يخفيها حسب السياسات.*

النتيجة النهائية:

بعد ما تعمل Scan بالكامل، هتعرف:

العنصر	المعنى
Data Stores	مصادر البيانات داخل BDE
Sensitive Data	الأعمدة اللي فيها بيانات حساسة
Proliferation	اتنقلت البيانات الحساسة منين لفين داخل الـ Mappings
Data Masking Verification	هل البيانات الحساسة اتحمت باستخدام DMO Transformation؟

مثال :

عندك Workflow في Big Data Edition:

- بياخد بيانات عملاء من HDFS
- يعديها على Mapping فيها DMO Masking
- يطلعها في Hive Table

هيحلل ده DPM 👀:

- يعرف إن عمود رقم البطاقة اتحرك من HDFS لـ Hive.
- ويتأكد إن الـ DMO فعلاً عمل Masking عليه.

بقيت الانواع نفس طريقة كتابة البرامتر

Jobs

يعني إيه "Job" في DPM؟

الـ **Job** يعني "مهمة أو عملية كبيرة" بيعملها النظام.

مثال بسيط:
أنت طلبت من DPM يعمل **Scan** على قاعدة بيانات فيها بيانات عملاء.

علشان ينفذ ده، DPM بيبدأ **Job** فيه خطوات كثيرة زي:

1. يتصل بالبيانات.
2. يقرأ الأعمدة.
3. يدور على بيانات حساسة.
4. يكتب النتائج.

ليه بنسميها Job؟

لأنها مش خطوة واحدة، دي **عملية فيها عدة مراحل**، والنظام بيتتبع كل خطوة فيها.

من إيه بيتكون الـ Job؟

كل Job بيتكون من:

- **خطوات Steps**: كل خطوة تنفذ حاجة (زي الاتصال بالبيانات - فحص الأعمدة - تصنيفها).
- **حالة Status**: كل Job له حالة زي (شغال، خلص، فشل...).

إمتى يتعمل الـ Job؟

- لما تعمل **Scan**.
- لما تستورد **Data Store**.
- لما تنفذ **سياسة حماية أو تصنيف**.
- لما تعمل **Export** أو **Import**.

أنواع الحالات اللي ممكن يكون فيها الـ Job:

الحالة	معناها
✅ Completed	المهمة خلصت بنجاح.
🔄 Running	شغالة دلوقتي.
⏸ Paused	متوقفة مؤقتًا.
❌ Failed	حصل فيها خطأ.

الحالة	معناها
Terminated	اتقفلت أو انت وقفتها بنفسك.

طيب أقدر أعمل إيه في الـ Job؟

تقدر:

- **توقفها مؤقتًا** (Pause)
- **تشغلها تاني** (Resume)
- **توقفها خالص** (Stop)
- **تمسحها أو تلغيها** (Terminate)

فين أتابع الـ Jobs؟

من صفحة اسمها **Jobs Workspace** في DPM:

- تشوف كل Job شغال أو خلص.
- تشوف لو فيه حاجة فشلت وليه.
- تقدر توقف أو تعيد التشغيل.

مثال:

تخيل عندك 3 قواعد بيانات:

- عملت Scan على أول قاعدة ← ده بقى اسمه Job رقم 1.
- عملت Scan على الثانية ← Job رقم 2.
- عملت Scan على الثالثة ← Job رقم 3.

كل Job منهم ليه خطوات، وحالة، وتقدر تتحكم فيه.