

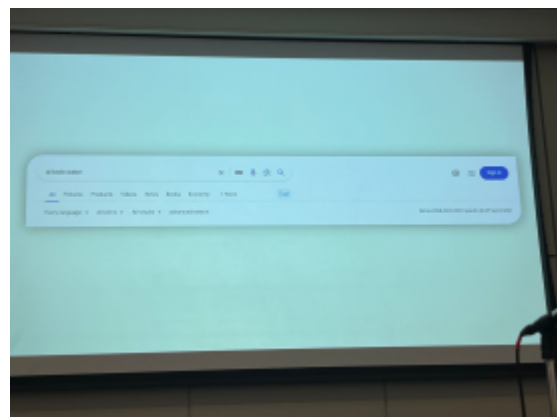
TS 1.1 How To Go Digital As A Water Utility

Cinter - A Human - Centric Data Management Platform
For The Water Sector, Ingemar Clementson, Sweden

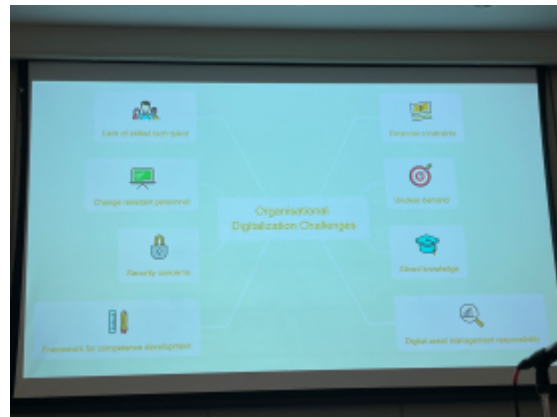
1. 발표 및 PPT 설명



발표자가 디지털화 전략가로 일하고 있는 NSVA는 스웨덴 남부에 위치한 공공기관으로, 빗물을 관리하고 하수 처리를 담당하며 음용수를 공급하고 있다. 발표 주제는 데이터 플랫폼 설계 작업이 프로젝트에서 작업한 내용을 바탕으로 3가지 주제에 대한 통찰 (문제점, 사람, 기술)이다.



인용구 - 여기서 빅데이터를 디지털화로 대체해도 말이 성립한다. AI 도구와 물 관련 검색을 하면 대략 9억 개의 답변이 나온다. 그런데 왜 우리는 이러한 도구들을 충분히 활용하고 있지 않는가?



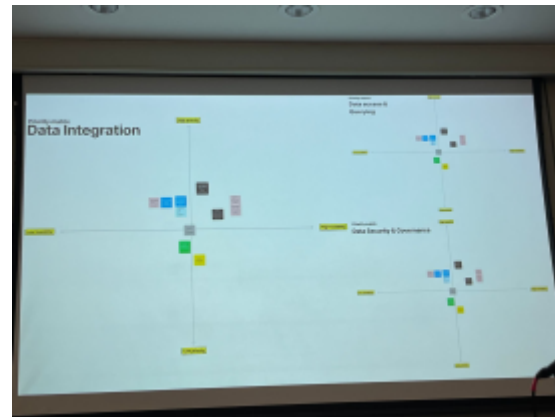
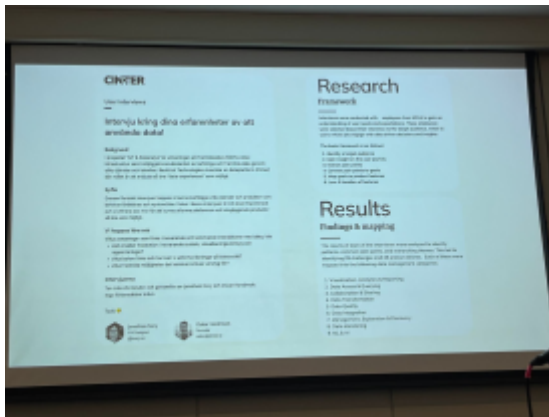
공공기관은 이러한 도구를 사용하여 비용을 절감하고 운영 효율성을 높일 수 있다. 하지만 이러한 기술들은 업계의 다양한 코너에만 구현되어 있으며, 비록 그 가치가 크지만 광범위하게 사용되고 있지는 않다. 이는 이러한 도구를 사용 가능하게 하기 위해 필요한 작업 방식의 변화에 대한 준비가 되어 있지 않기 때문이다.

조직 디지털화에 필요한 과제 1. 일반적으로 훨씬 높은 수준의 기술 필요 2. 일하는 방식의 변화를 요구 3. 투자를 필요로 하며, 4. 지식이 직원에서 시스템으로 이전 필요 5. 매우 견고한 디지털 자산 관리 필요 많은 데이터 소스와 데이터 유형을 관리할 때, 데이터는 내부적으로 누가 소유하고, 품질에 대한 책임은 누구에게 있으며, 정보는 누구에게 전달되고, 가장 중요한 정보와 조직으로서 실제 어떤 정보에 대해 행동을 취하는지에 대한 체계가 잡혀있어야 한다.



그래서 이러한 문제를 해결하기 위한 프로젝트 : Cinter

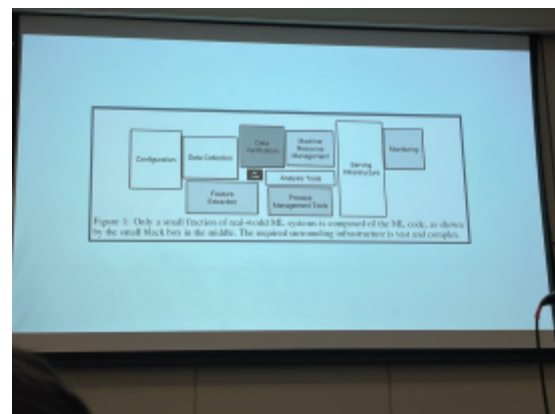
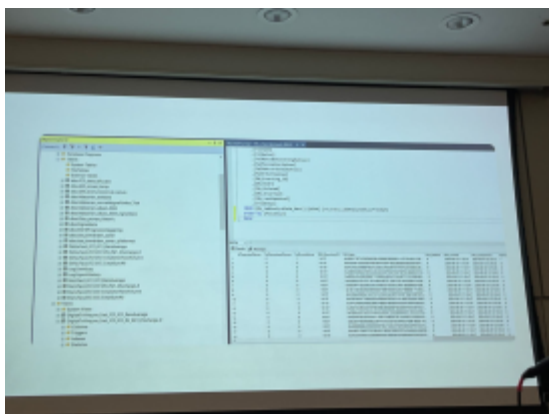
NSVA와 Sweden water research는 수십억 개의 데이터와, 수백만 개의 IoT 장치를 처리할 수 있는 시스템을 구축할 기술적 역량이 없다는 것을 깨닫고 위와 같은 협력 업체를 찾았다.



데이터 사용 경험에 대한 인터뷰 - 직장 동료공동 목표를 가진 프레임워크 내에서, 데이터 사용 경험에 관한 다양한 답변을 받음.데이터 엔지니어링 관점에서, 많은 질문이데이터가 잘 못된 것인지,측정이 잘못된 것인지,단순한 시스템 변동인지판단하기 어렵다는 점에 초점을 맞추고 있음.

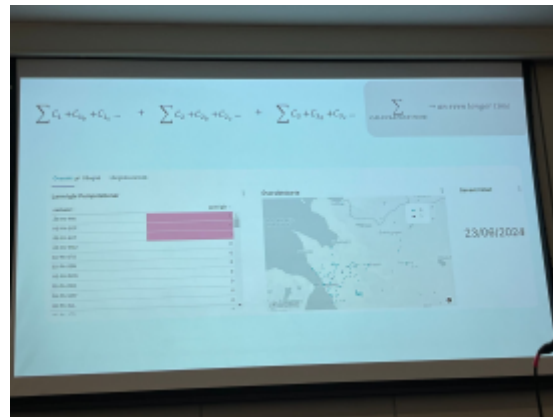
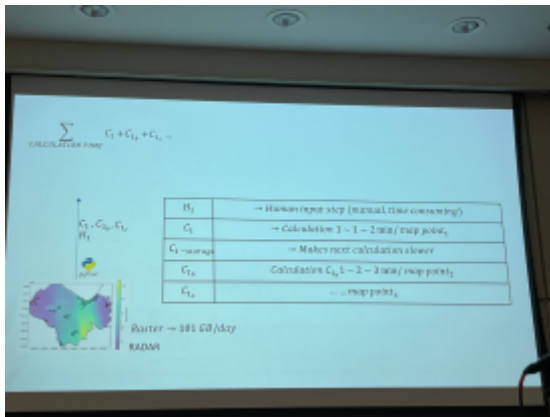
모든 답변을 priority matrix에 넣어서,무엇이 높은 우선순위인지,낮은 우선순위인지,무엇이 실현 가능한지, 불가능한지 파악함.여기에 포함할 수 없는 답변도 많았지만, 이는 후속 작업을 위한 견고한 기초가 되었음.

답변을 요약하자면, 기술적으로 뛰어난 많은 물 엔지니어들은 데이터를 저장하는 시스템에 들어가 작업하는 것을 두려워하며, 무언가를 삭제하는 것과 실험하는 것을 두려워한다. 또한 데이터를 사용하여 코딩을 배우고 사용하는 동안 다른 동료와 협력하기가 매우 어렵다.



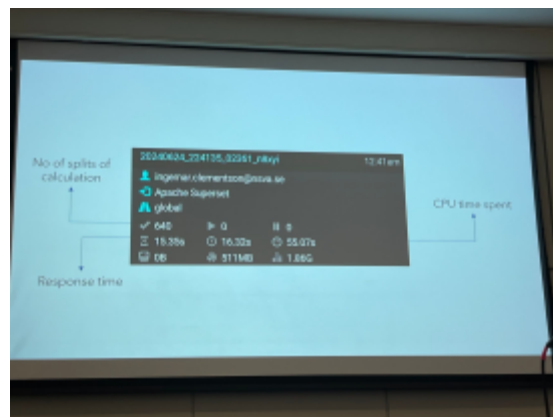
Cinter의 데이터 카탈로그, 데이터를 쉽게 관리하고 주석을 달 수 있으며, 쉽게 공유할 수 있음또한, 데이터 사용 현황을 추적하며, 이는 사람들을 감시하기 위함이 아니라 회사 내에서의 역할에 따라 데이터 사용에 대한 제안을 제공하기 위함이다.

이 그림은 머신 러닝 프로젝트에서 사용된 시간을 정량화한 과학 논문에서 가져온 그림.가운데 검은 상자는 실제 머신 러닝 코딩에 사용된 시간으로, 우리가 해야 할 모든 작업은 단순히 작동되어야 한다는 것이다.



예시) 날씨 레이더를 사용하여 강우강도와 지속시간을 계산하고 싶고, 날씨 예보를 통해 다음에 올 비를 예측하고 싶을 때, 프로그래밍 언어로 Python을 사용한다면, 계산 하나 하나에 사람의 입력 시간, 다음 계산, 저장 등으로 시간이 많이 소요된다.

또한, SQL을 활용하여 데이터를 추가할 때나 온라인 API에서 데이터를 결합할 때에도 많은 시간이 소요된다. 또한 이를 여러 곳에서 사용 가능한 대시보드에 포함하고 싶으며, 이 시스템이 몇 분마다 업데이트되기를 원한다면 더욱 많은 시간이 소요된다.



이 시스템은 사용자가 자신의 저장 플랫폼을 선택할 수 있도록 설계되었으며, 자신의 계산 엔진을 선택할 수 있다. 또한 특별한 보안 규정을 설정할 수 있고, 계산 및 저장시 가장 저렴하고 효율적인 공급업체를 선택하여, 계산 비용을 절감할 수 있다. 이 프로젝트는 Iceberg라는 product도 존재하며, 이는 data lake의 질서를 부여하는 데 사용되며, 데이터를 추출하고 조직화하는 속도를 높여준다. Trino는 프로그래밍 언어의 orchestrator로 사용되며, 이는 계산 계획을 관리하고, 계산을 다양한 부분으로 분할하여 더 효율적으로 작동할 수 있도록한다. 또한 다른 소프트웨어와 연결하여 데이터를 사용할 수 있도록한다. 이는 새로운 제품을 구현하는 문턱을 낮추며, 정제된 데이터를 Trino를 통해 전송할 수 있다. 데이터 카탈로그는 Jupyter 노트북을 통해 데이터에 접근할 수 있으며, 오픈 소스 BI 도구인 Superset도 사용한다. 여기에 사용된 모든 제품은 오픈 소스 제품으로, 필요에 따라 구성요소를 변경할 수 있다. 이 시스템은 더 많은 오픈 소스 소프트웨어를 사용하고 있으며, 단순히 분산된 쿼리를 사용한 것이다.

이것이 15초 만에 응답한 큰 쿼리이다. 이 엔진은 640개의 부분으로 나뉘어 계산되며, 실제 소요 시간은 단일 컴퓨터 기준 55초이다. 그리고 쿼리가 커질수록, 이 시스템의 장점은 커지며, 여러 곳에 동일한 질문을 하고자 할 때 도움이 된다.

Why is big data so hard?

Velocity

The speed at which data is generated

Volume

The amount of data from myriad sources

Variety

The types of data; structured, semistructured, unstructured

빅데이터가 어려운 이유 velocity, volume, variety 모두 중요함. 사용해야 하는 기술과 기술이 설계되는 방식에 대한 요구가 발생하며, 이로 인해 사람들이 일하는 방식의 변화가 필요함. 디지털화와 기술 개발이 IT나 개발에만 국한된 것이 아니라, 경영과 리더십의 문제이다.

2. 요약 및 정리

1. AI와 디지털화의 중요성

- AI 도구와 디지털화를 활용하여 공공 기관의 비용을 절감하고 운영 효율성을 증대시킬 수 있다.
- 그러나 물 산업 전반에 AI 구현이 널리 이루어지지 않았으며, 이는 조직이 준비되지 않았고 높은 기술 요구 사항이 있었기 때문이다.
- 이러한 기술 도입은 조직 구조와 업무 방식을 변화시켜야 했다.

2. 데이터 플랫폼 설계와 초기 단계

- NSVA와 Sweden Water Research는 대규모 데이터를 처리할 역량이 부족하여, 파트너십을 통해 기술적 역량을 보완했다.
- 내부 설문조사를 통해 조직 내 데이터 사용 경험을 조사하고, 이를 바탕으로 데이터 플랫폼 설계에 필요한 요구 사항을 도출했다.
- 데이터 관리와 협업에 대한 어려움이 주요 문제로 확인되었으며, 이를 해결하기 위해 데이터 카탈로그를 설계했다.

3. 기술적 구현

- 데이터 카탈로그를 통해 데이터를 쉽게 관리하고 주석을 달며, 사용 현황을 추적할 수 있도록 했다.
- 시스템은 사용자에게 맞춤형 저장 플랫폼과 계산 엔진을 선택할 수 있도록 유연하게 설계되었다.
- 데이터 레이크 하우스를 통한 데이터 구조화와 Trino를 통한 효율적인 계산 관리로 시스템의 성능을 극대화했다.

4. 빅 데이터의 도전 과제

- 물 산업에서 빅 데이터의 복잡성과 다양한 데이터 형식이 큰 도전 과제였다.
- 기술 개발이 조직 운영 방식의 변화와 밀접하게 연관되어 있으며, 이러한 변화가 기술의 성공적인 도입에 필수적이었다.

5. 결론

- 디지털화와 기술 개발은 IT나 개발 부서의 문제가 아니라 경영과 리더십의 문제로서, 조직 전체의 협력이 필요하다.