NIWA

Taihoro Nukurangi

Climate, Freshwater & Ocean Science

NIWA Capabilities 2025

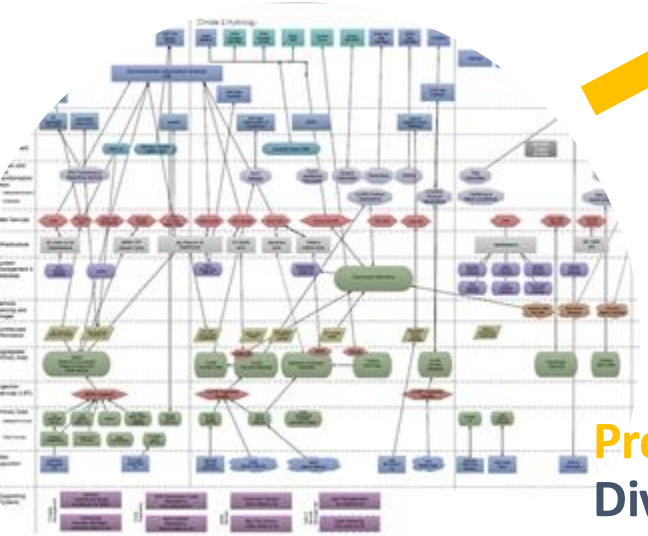
# NIWA Data Transformations

*Jochen Schmidt, Chief Scientist, Environmental Information*

# Data
## Transformation


**Future:**
**Data Integration**

**Current:**
**Siloed Solutions**



**Pre 2010:**
**Diverse Systems**

# Data Transformation

## Towards an Integrated Data Platform

# Vision

NIWA Statement of Corporate Intent:

"Develop an integrated **Data Platform** that improves data management and access and seamlessly enables data science / analytics to derive new insights from various data sources"

# Transformation through the Integrated Data Platform

**Current**

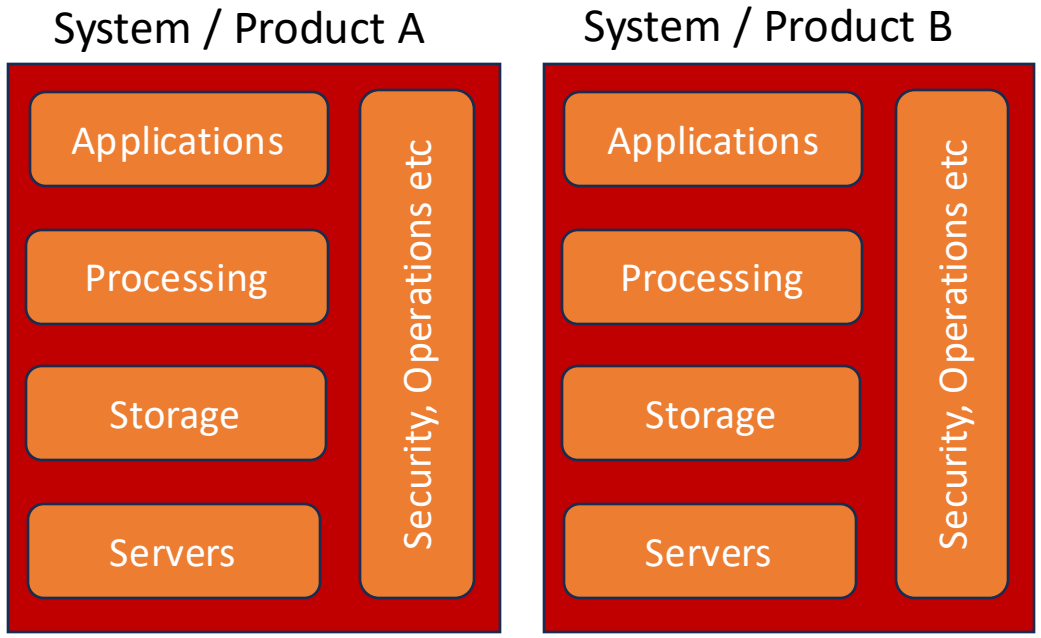
- Data 'silos' with nothing shared
- Applications are a small part of overall operations mixed in with things that don't add customer value
- We build everything for every new product/system

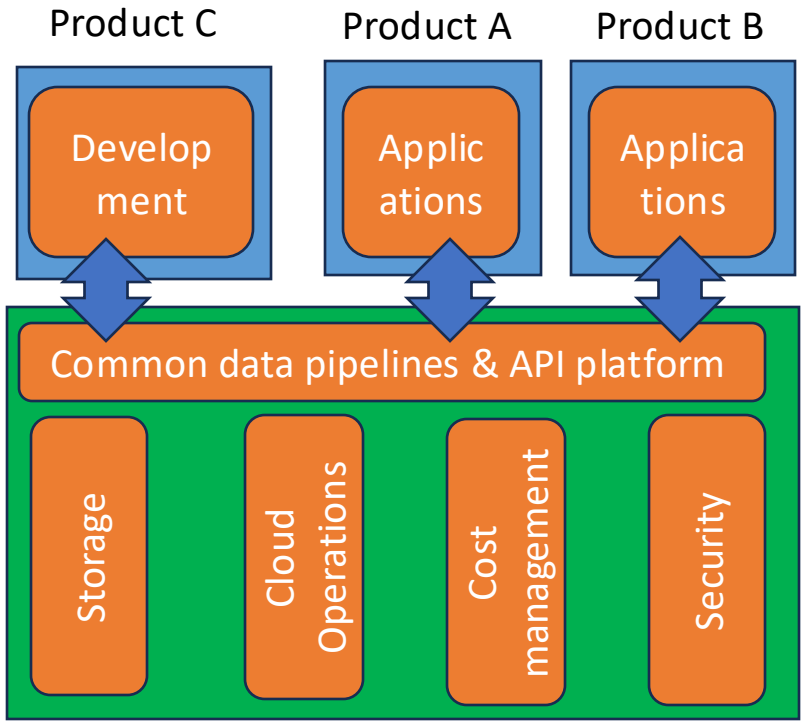
**-> Stuck support legacy systems, no space to do new things**

**Future**

- Common Platform enables sharing of benefits (data, code)
- Focus on new products and continued improvements
- Support high number of data product (still have flexibility to choose from cloud products)

**-> Lower ongoing cost, ability to create new things over time**



# Outcomes

- **Integrated** / unified data access
- Ready for more diverse data (types, sources) & scalable
- Common technologies & patterns for shared development and re-use
- **Automation** to the max
- **'AI enabled'**

# Data Transformation: Impact / Benefits

One.NIWA
Governance
Sovereignty
Security

Organisational visibility through enterprise approach

‘Big Data’ ready scalable

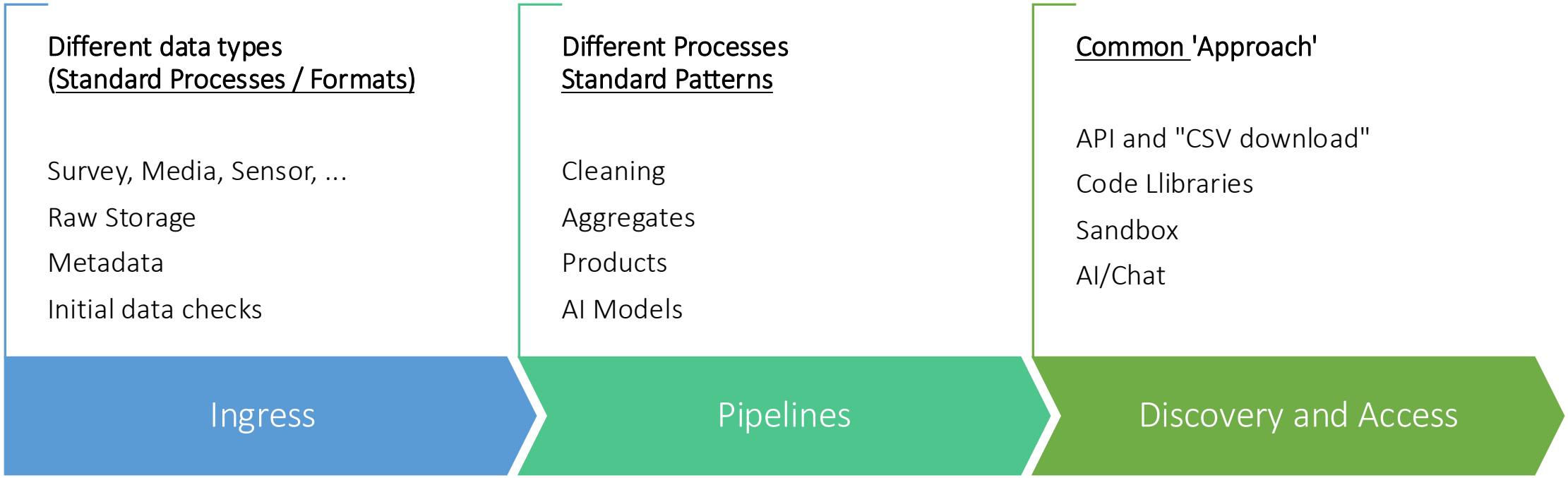
Consistent
Storage
Pipelines
Processing

Automation built in

Data Science
Innovation
Experimentation
Sandbox
built In

Modern
technologies
SaaS

# High level architecture



# Data Platform Technologies


**Different data types (Standard Processes / Formats)**

Survey, Media, Sensor, ...

Raw Storage

Creates Metadata

Initial data checks

**Different Processes / Standard Patterns**

Cleaning

Aggregates

Products

**Common 'Approach'**

API and "CSV"

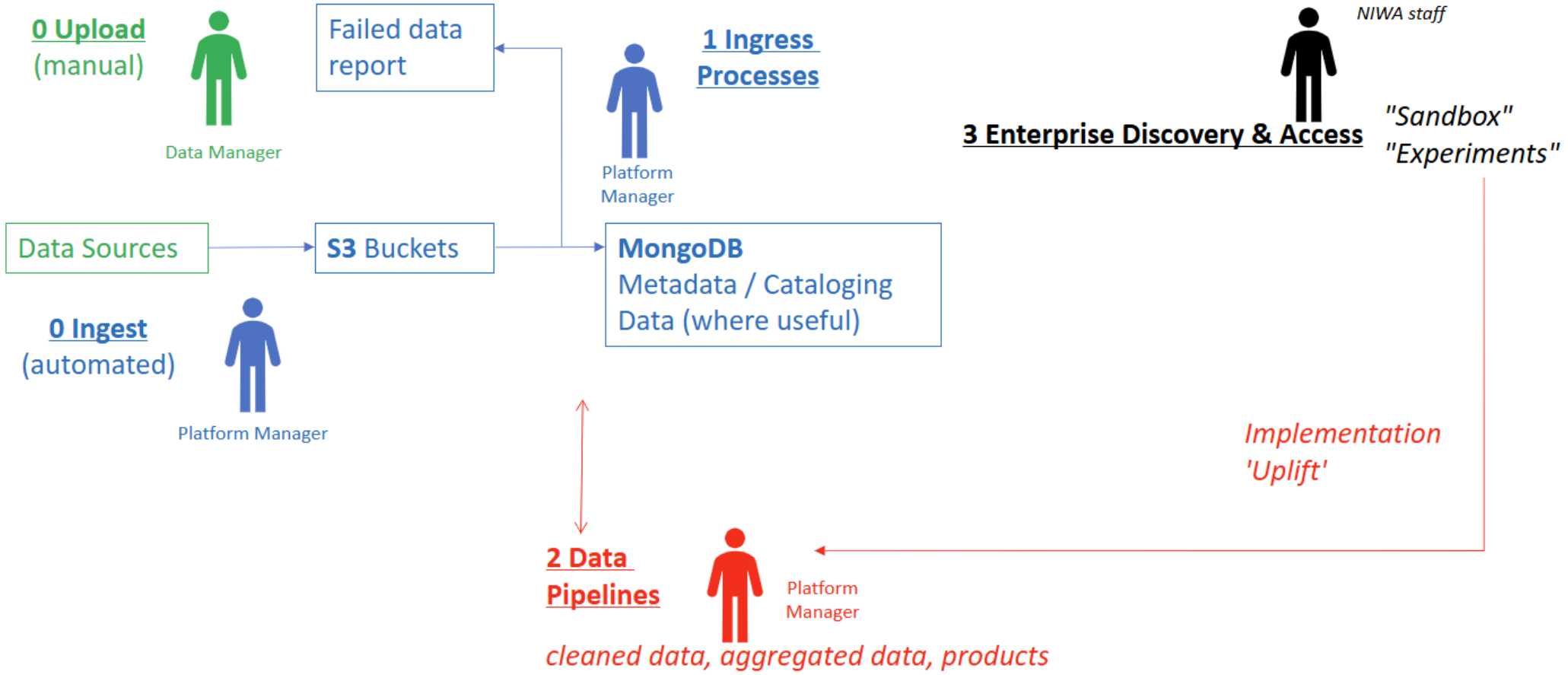
Llibraries

Ingress

Pipelines

Discovery and Access

# Data Platform Implementation Patterns

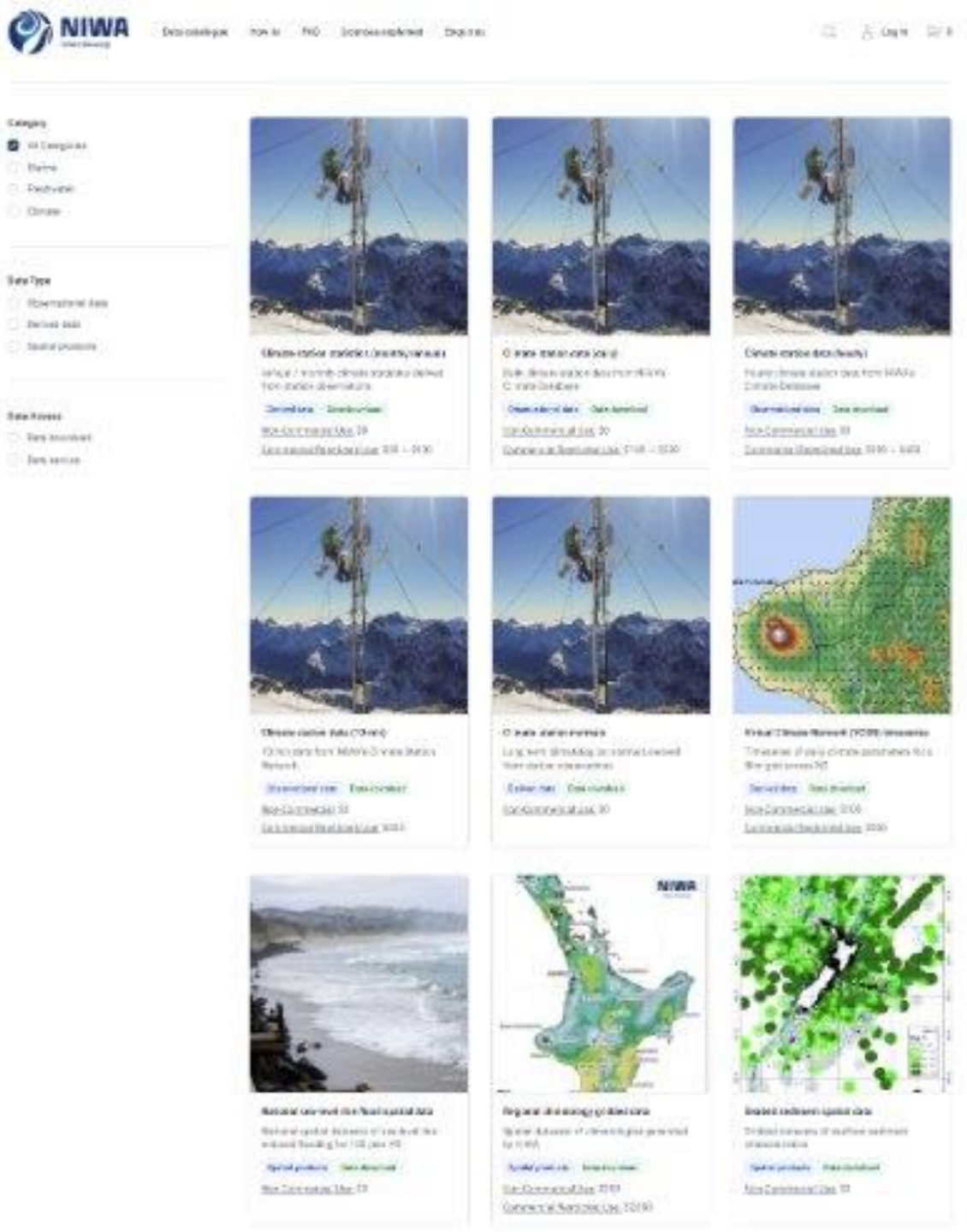


# Example 1: DataHub

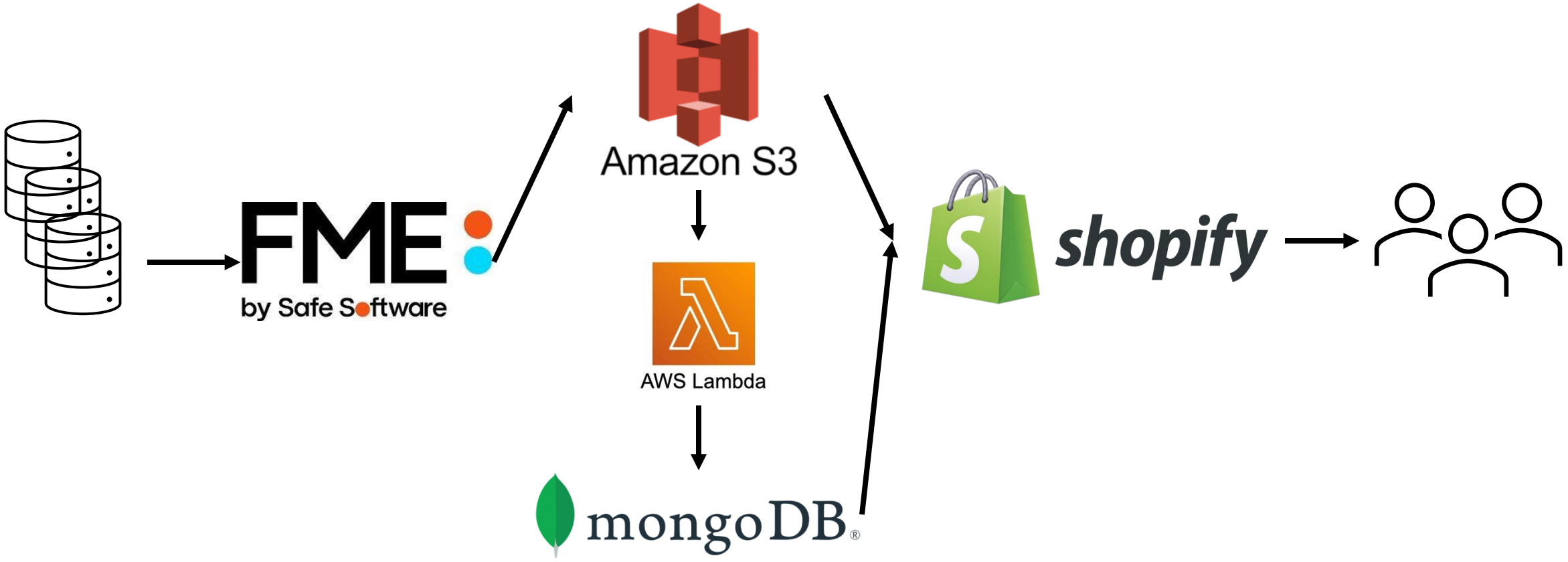
Integration of various data and making it available for download

# DataHub: What is it?

- "One Platform" for simple public download access to limited datasets
- Data can be stored and provided as required (e.g. format)
- Easy configurable / add new products
- Standard Licenses, Limit access & Credit Card payment



# DataHub Technologies & Architecture



# DataHub Processes

New product

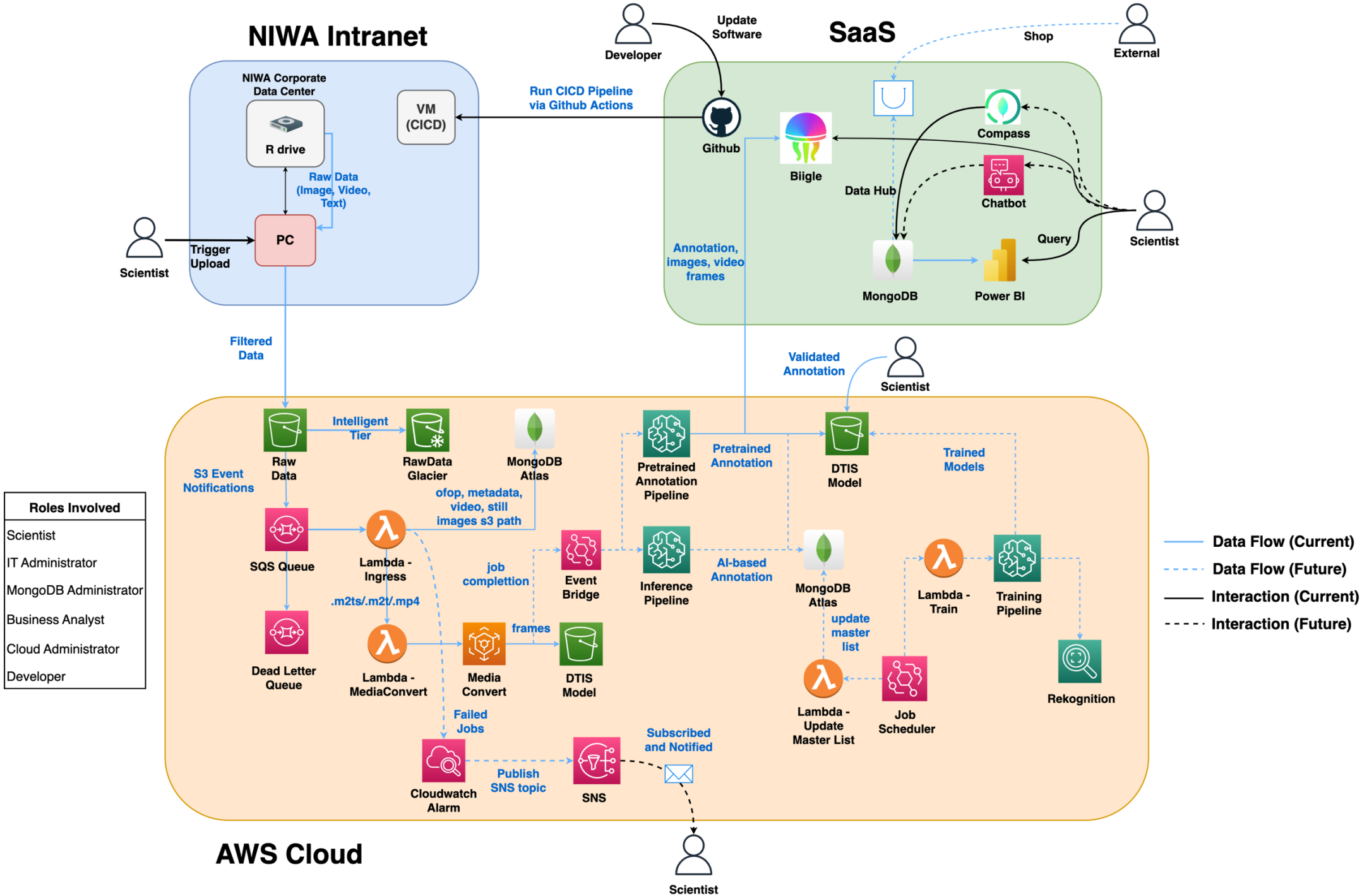
- Shopify configuration
  - Web content
  - Business logic & Price
- Data ingest process
  - S3 bucket
  - FME Workflow
- Automation
  - Metadata sync to MongoDB
- Held together by conventions
  - Product ID
  - Standard data product fields

BAU Functionality “Out of the box”

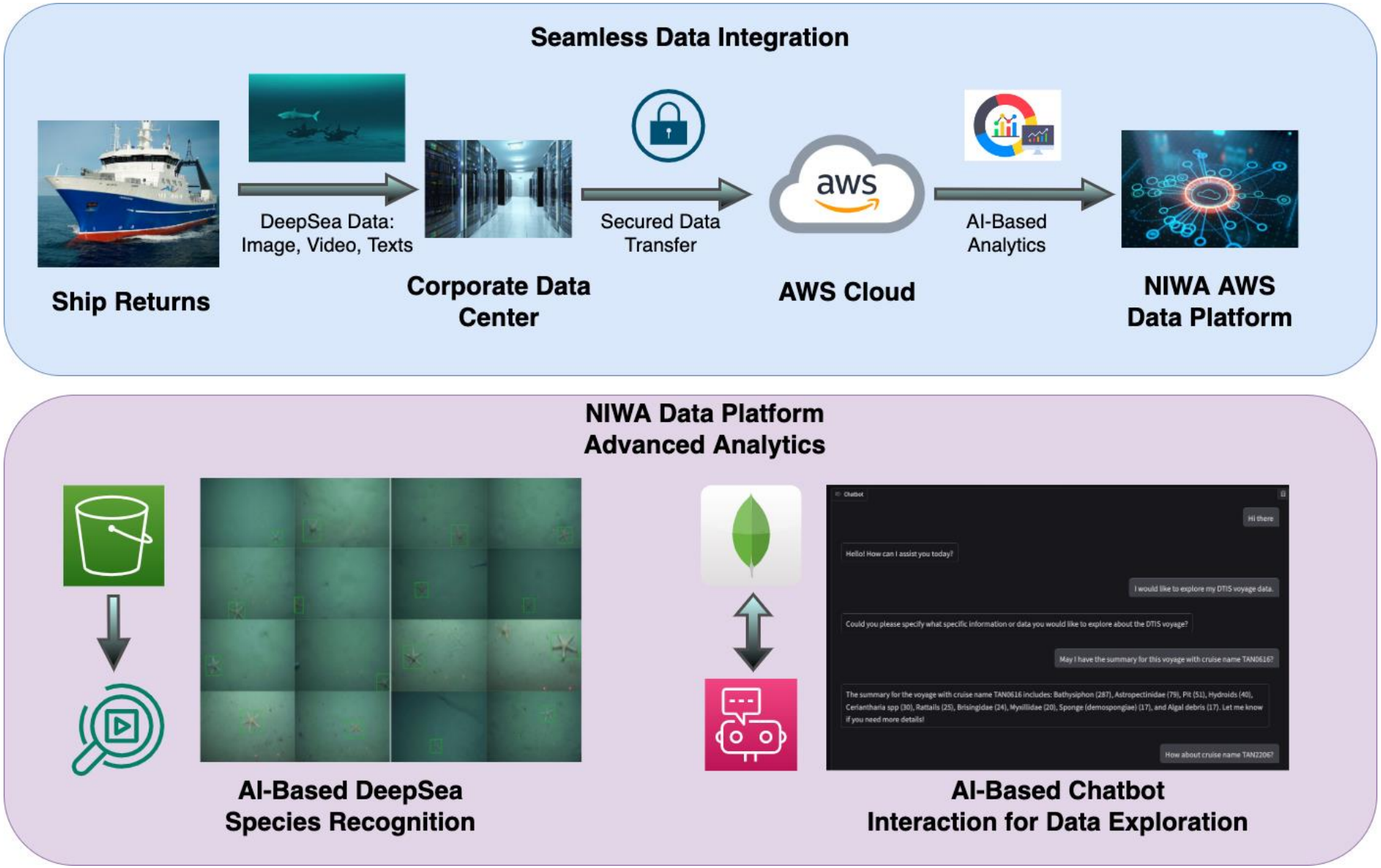
- Customer Management
- Order Management
- Discounts
- Analytics / Reports

# Example 2: Data Pipeline

- Integration of poorly managed data and AI enabling
- Sea floor Imagery / Videos + Manual labels
- Decades of data


NIWA Intranet
NIWA Corporate Data Center
R drive
VM (CICD)
Raw Data (Image, Video, Text)
PC
Scientist
Trigger Upload
Run CICD Pipeline via Github Actions
Developer
Update Software
SaaS
Github
Biigle
Shop
External
Compass
Chatbot
Data Hub
Scientist
Query
MongoDB
Power BI
Annotation, images, video frames
Filtered Data
Validated Annotation
Scientist
Raw Data
Intelligent Tier
RawData Glacier
MongoDB Atlas
S3 Event Notifications
ofop, metadata, video, still images s3 path
Pretrained Annotation Pipeline
Pretrained Annotation
DTIS Model
Trained Models
SQS Queue
Lambda - Ingress
.m2ts/.m2t/.mp4
job completion
Event Bridge
Inference Pipeline
AI-based Annotation
MongoDB Atlas
Lambda - Train
Training Pipeline
Dead Letter Queue
Lambda - MediaConvert
Media Convert
frames
DTIS Model
update master list
Lambda - Update Master List
Job Scheduler
Rekognition
Failed Jobs
Cloudwatch Alarm
Publish SNS topic
SNS
Subscribed and Notified
Scientist
AWS Cloud
Roles Involved
Scientist
IT Administrator
MongoDB Administrator
Business Analyst
Cloud Administrator
Developer
Data Flow (Current)
Data Flow (Future)
Interaction (Current)
Interaction (Future)

# Data Platform DTIS Use Case



# The Future

Value creation through

- Bringing together diverse big data sources
- ‘Unlocking the potential of AI’
- Integration with HPC

NIWA

Taihoro Nukurangi

Climate, Freshwater & Ocean Science