2019 USENIX Conference on Operational Machine Learning

NTT DATA
Trusted Global Innovator



A Distributed Machine Learning For Giant Hogweed Eradication

Naoto Umemori and Masaru Dobashi

May 20, 2019

OSS Professional Services, NTT DATA Corporation

© 2019 NTT DATA Corporation

Agenda

1. Introduction
2. Our OpML Experience - Giant Hogweed Eradication -
3. Lessons Learned while Adopting ML in SYS

© 2019 NTT DATA Corporation

Introduction



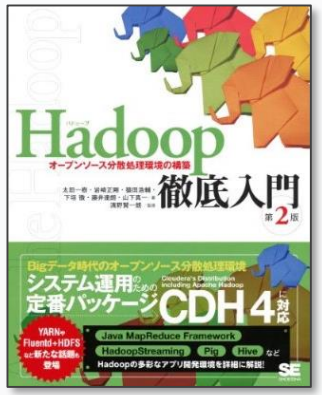
© 2019 NTT DATA Corporation

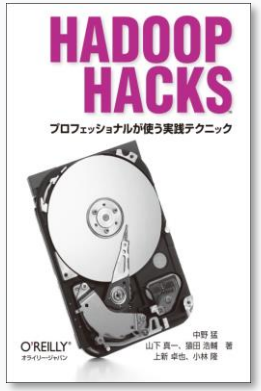
Our Capabilities

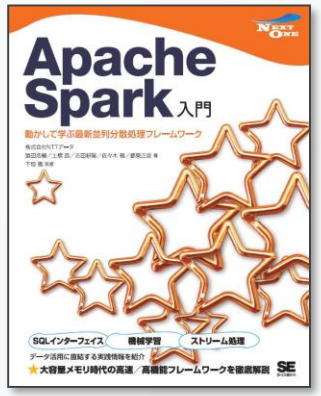
Expert / Professional Team of Open-Source Software for 15+ Years

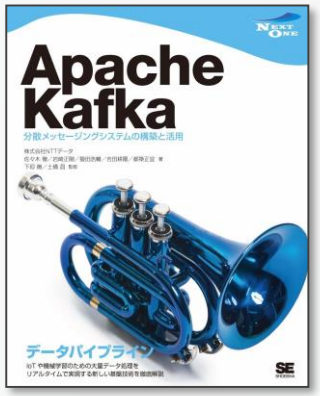
Hadoop/Spark/Kafka Capabilities

- World's no.7 in the Big Data era led by Hadoop. (as of 2017)
- Consulting, Design, Deploy and Operation of Hadoop Clusters, and Publication









These books were written by our team members.

Experiences

- 10+ years of experience (in Distributed Computing)
- 100+ production cases (Hadoop Clusters in the range of 10 - 1200+ nodes)
- Solution that covers security, application development, cluster construction etc.
- Customers in a wide range of industries (Automotive, Enterprise, Financial, Telco, etc.)

© 2019 NTT DATA Corporation

Our OpML Experience - Giant Hogweed Eradication -



© 2019 NTT DATA Corporation

NTT DATA

Overview of Giant Hogweed Eradication Project

Automating the detection of highly toxic plants by exploiting image recognition/detection processing on a distributed processing platform

Collect

4K IMAGES



Data Sources

Store

TensorFlow™

APACHE Spark™

APACHE hadoop

Data Lake

Utilize

OpenStreetMap

SAP HANA



Data Analytics

© 2019 NTT DATA Corporation

NTT DATA

Challenges of The Project

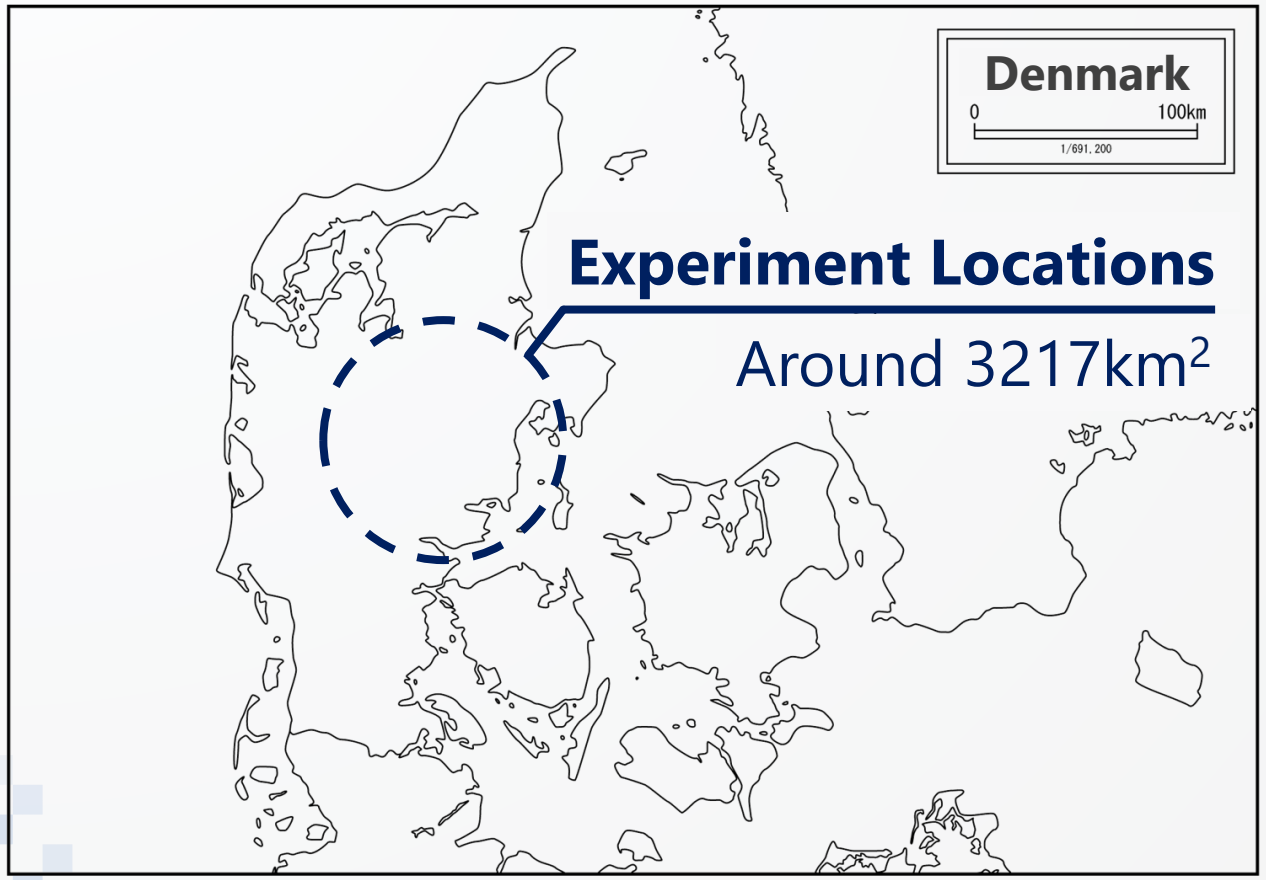


1. Data Volumes
2. Preparation of Supervised Data
3. Coordinate Calculation at Pinpoint

© 2019 NTT DATA Corporation

Challenge.1: Data Volumes

200.0+ Terabyte



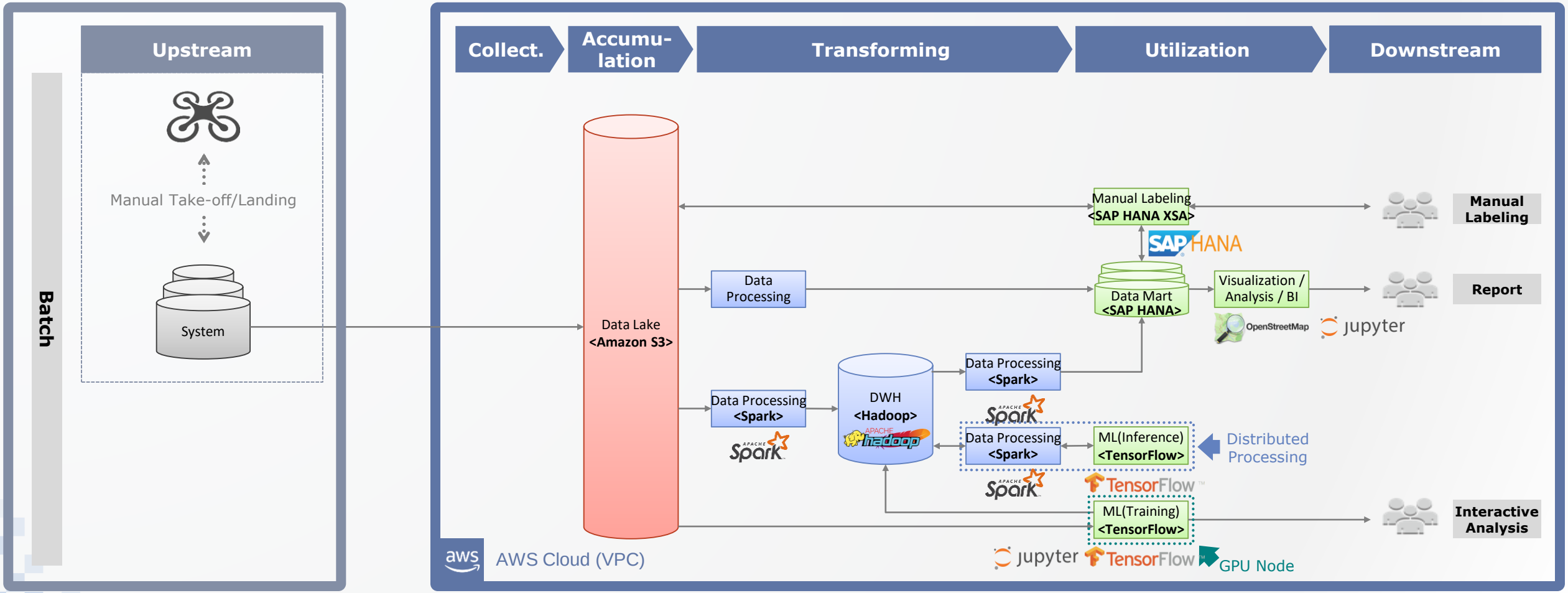
Parameters	Values
Farmland rate of Denmark*1	62.01 [%]
Estimated value of land area of Experiment Location	1,994.86 -> 2,000 [km²]
Size of an image taken by Drone	10 [MiB]
Aspect ratio of an image	4 : 3
Drone aviation altitude	20 [m]
FOV: Field of View	90 [Degree]
The estimated number of Images	20,000,000
Estimated data volume	**200.0 [TiB]**

*1: https://ecodb.net/country/DK/nature/

© 2019 NTT DATA Corporation

Our Approach and Solution.1

- **Preparation:** SAP HANA is used to prepare Supervised Data
- **Machine Learning: [Training]** TensorFlow is running on a single GPU node.
  [Inference] TensorFlow is running on Hadoop/Spark Cluster. (Distributed Processing)



© 2019 NTT DATA Corporation

Challenge.2: Preparation of Supervised Data

Preparation of Supervised Data is a hard task. It is more difficult to identify Giant Hogweed from aerial images.

- Lack of supervised data for Training is a common story in ML.
- However, there are cases where it is difficult even to prepare what can be correct data. For instance, identifying Giant Hogweed from an aerial image of Drone needs specialized knowledge by biologists.



Giant Hogweed (Positive Class)



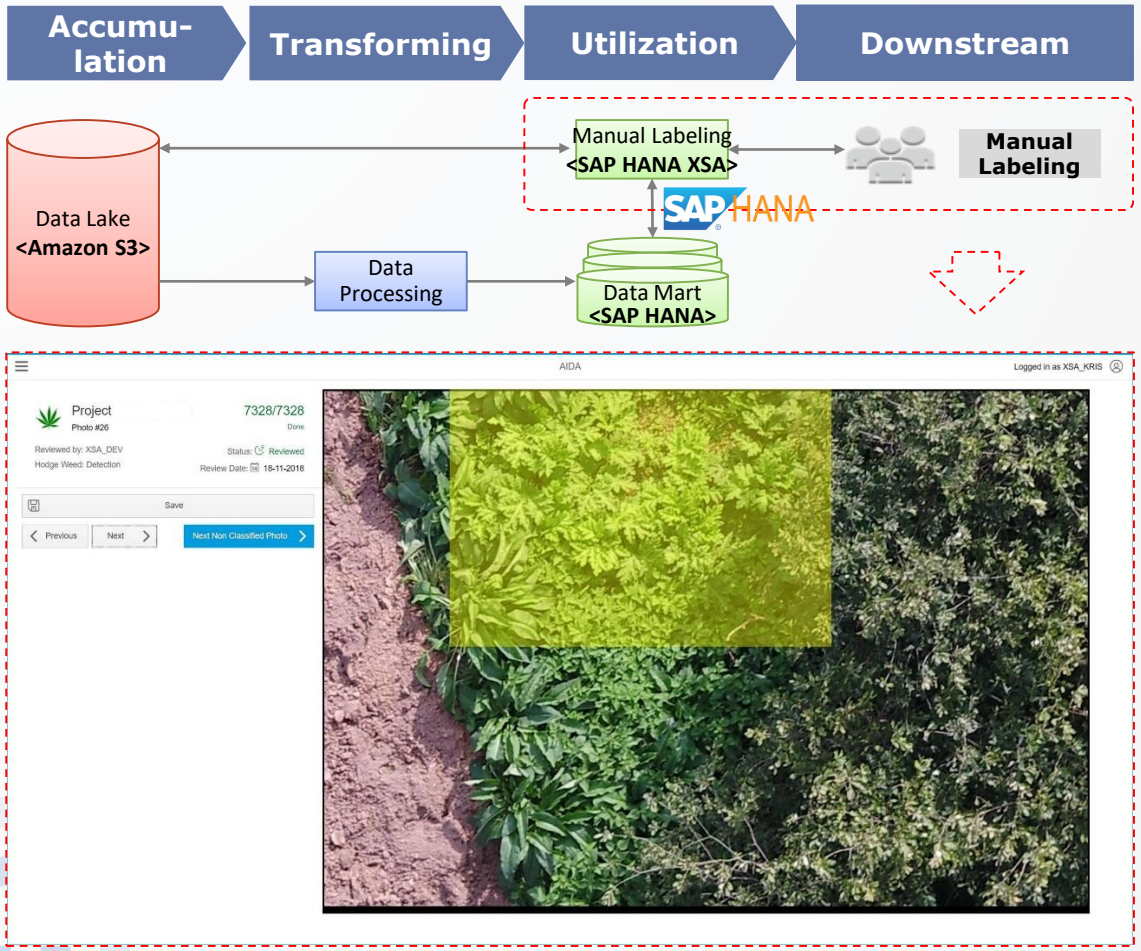
Weed - NOT Giant Hogweed (Negative Class)

Could you tell the Giant Hogweed from Weed? :(

© 2019 NTT DATA Corporation

Our Approach and Solution.2

Introduce the Labeling Application for Data Preparation

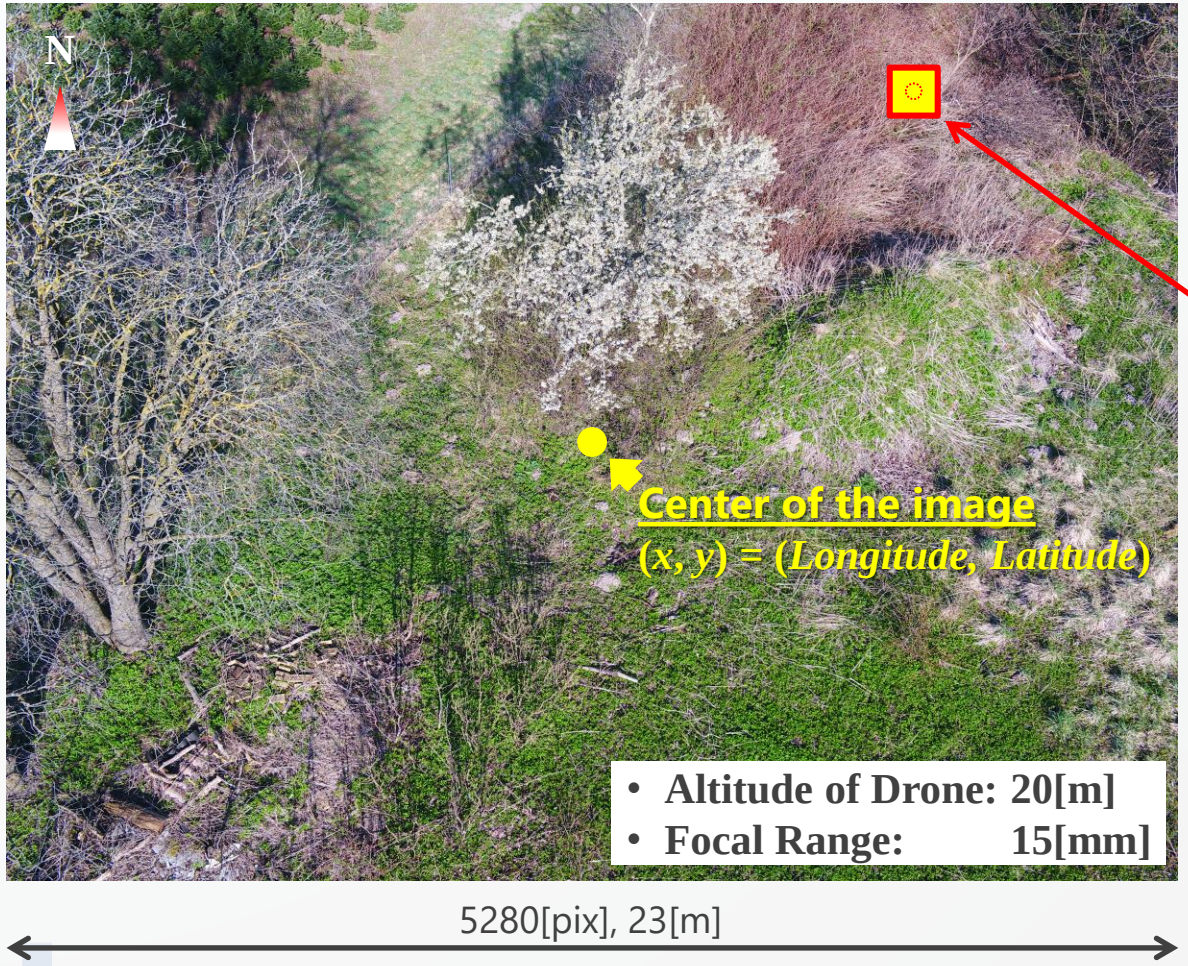


- A user-friendly UI that non-engineers can operate intuitively.
- Prepare a slightly larger image and label it by an expert on a small 6x4 split image.
- Labeled data is stored as supervised data in Data Lake.
- Note: There is no correction mechanism for what is labeled

© 2019 NTT DATA Corporation and itelligence

Challenge.3: Coordinate Calculation at Pinpoint

No Coordinate Calculation at Pinpoint, No Giant Hogweed Eradication



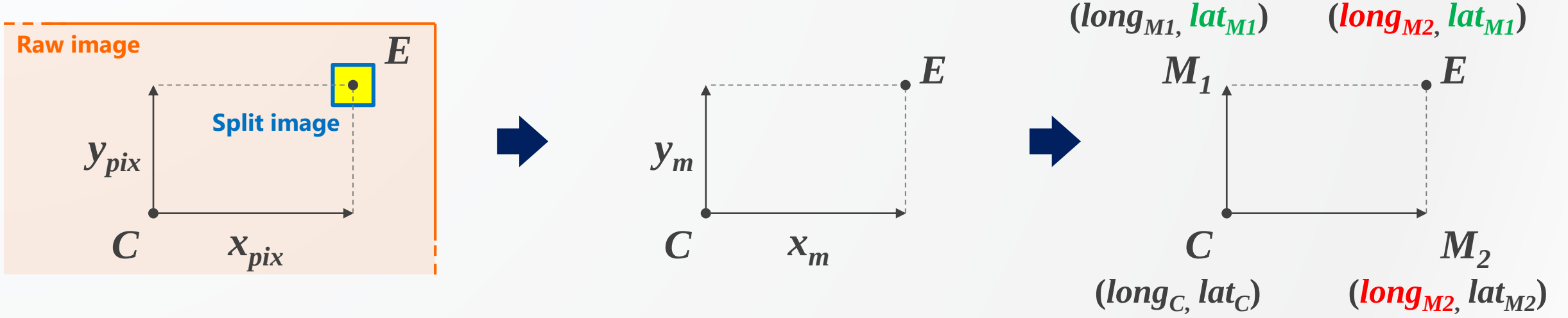
- The aerial image range taken by drone is quite wide.
- Images taken aerially with Drone have latitude and longitude information. However, **the input 24x18 split image(220x220 pix)** of inference processing is obtained by dividing an aerial image, and **the image does not have latitude and longitude information.**
- In order to eradicate Giant Hogweed, it is necessary to estimate the latitude and longitude of the habitat by pinpointing.

© 2019 NTT DATA Corporation

Our Approach and Solution.3

Approximated geo coordinate calculation

The spherical plane of the earth can be approximated to the horizontal plane because the length of on a side is short (0.96[m], 220[pix]).



Calculate the distance and direction from the center coordinates of raw image(C) to the center coordinates of split image(E).

Convert pixel values to physical distances by using exif.

Calculate the latitude and longitude of the split image by approximating the earth ellipse to horizontal plane.

© 2019 NTT DATA Corporation

Summary up to here



1 Data Volumes

- Distributed processing based on Apache Hadoop/Spark

2 Preparation of Supervised Data

- Labeling Application for Data Preparation based on SAP HANA Platform

3 Coordinate Calculation at Pinpoint

- Approximated geo coordination calculation

© 2019 NTT DATA Corporation

Lessons Learned while Adopting ML in SYS



© 2019 NTT DATA Corporation

NTT DATA

Context of Our Architecture Design



ML Systems

- Configuration
- Data Collection
- Data Verification
- Machine Resource Management
- Serving Infrastructure
- Monitoring
- Analysis Tools
- Feature Extraction
- Process Management Tools
- ML Code

Scaled ML Systems

- Orchestration
- Data Pipelines
- Data Verification
- System Resource Management
- Distributed-Serving Infrastructure
- System Monitoring
- Analysis Platform
- Feature Extraction
- Job Management Platform
- Distributed-ML Code

From "Hidden Technical Debt in Machine Learning Systems", D. Sculley at al. (Google), paper at NIPS 2015

© 2019 NTT DATA Corporation

Context of Our Architecture Design



ML Systems

- Configuration
- Data Collection
- Data Verification
- Machine Resource Management
- Serving Infrastructure
- Monitoring
- Analysis Tools
- Feature Extraction
- Process Management Tools
- ML Code

Scaled ML Systems

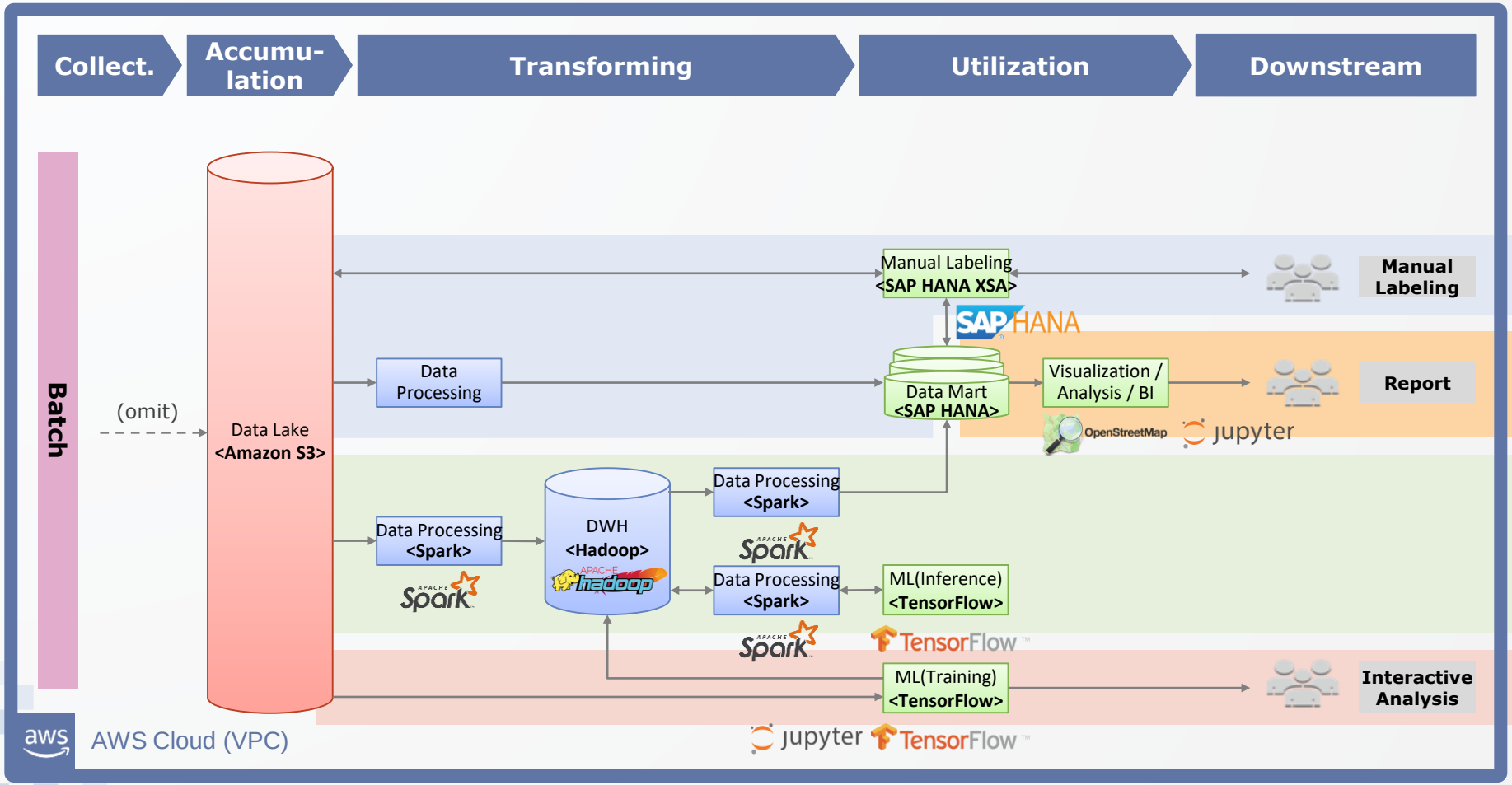
- Orchestration
- Data Pipelines
- Data Verification
- System Resource Management
- Distributed-Serving Infrastructure
- System Monitoring
- Analysis Platform
- Feature Extraction
- Job Management Platform
- Distributed-ML Code

From "Hidden Technical Debt in Machine Learning Systems", D. Sculley at al. (Google), paper at NIPS 2015

© 2019 NTT DATA Corporation

Data Pipelines for Scalable ML Systems

Inference: Hadoop/Spark + TensorFlow (Each Spark job natively calls TensorFlow libs.)
Training: (Future work)



Preparation pipeline

Analysis pipeline

Inference pipeline

Training pipeline

© 2019 NTT DATA Corporation

Distributed-ML Code



There are four approaches for Distributed-ML Model Development and Model Operation

Approach: Dev 2 Op	Development	Model Workflow	Operation
1: ML-friendly to ML-friendly	ex. python, TensorFlow	AP modification	Distributed TF, Custom FW
2: ML-friendly to SYS-friendly	ex. python, TensorFlow	Export → Model → Import	TFonSpark, BigDL
3: SYS-friendly to ML-friendly	N/A		
4: SYS-friendly to SYS-friendly	BigDL, DL4J	Scale out	BigDL, DL4J

We have chosen approach-2.

© 2019 NTT DATA Corporation

NTT DATA

LL: Integration of Apache Spark and TF

The complexity of processing logic changes depend on whether the inference processing implementation is Spark API based or the TensorFlow API based.

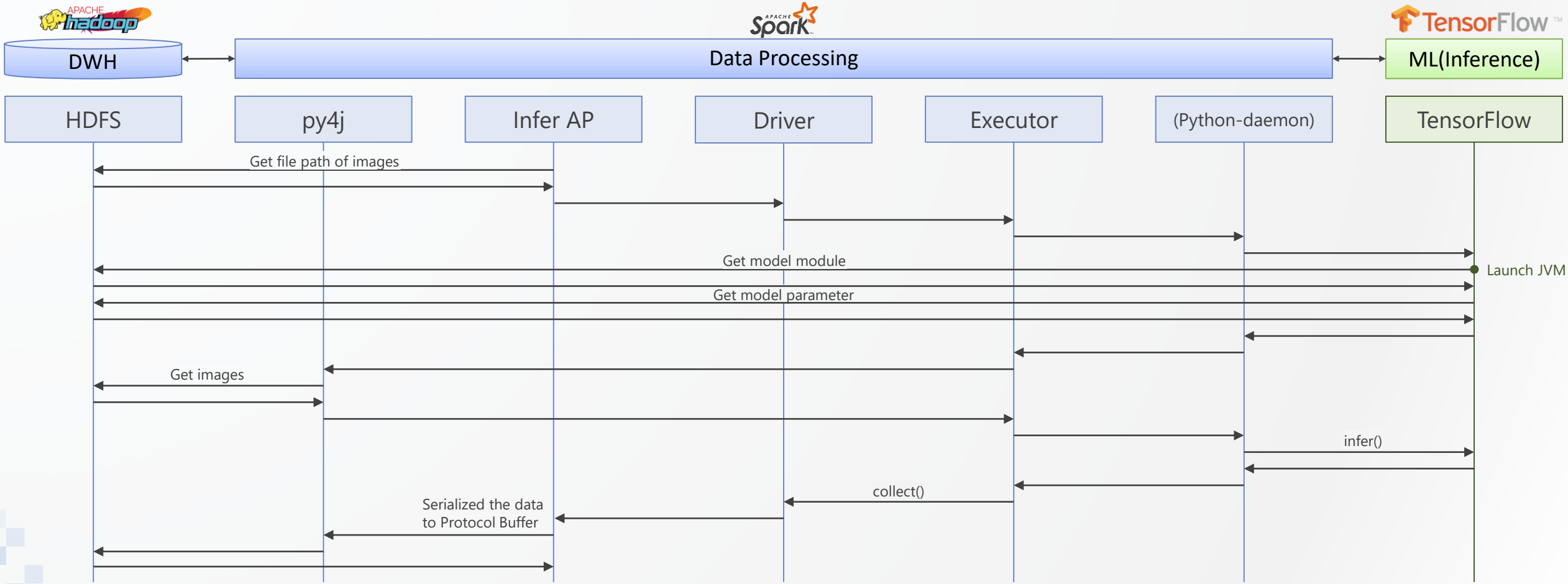


Figure: Sequence Diagram in case of **TensorFlow API-based** inference processing **w/o HDFS command** (25)

© 2019 NTT DATA Corporation

LL: Integration of Apache Spark and TF

The complexity of processing logic changes depend on whether the inference processing implementation is Spark API based or the TensorFlow API based.

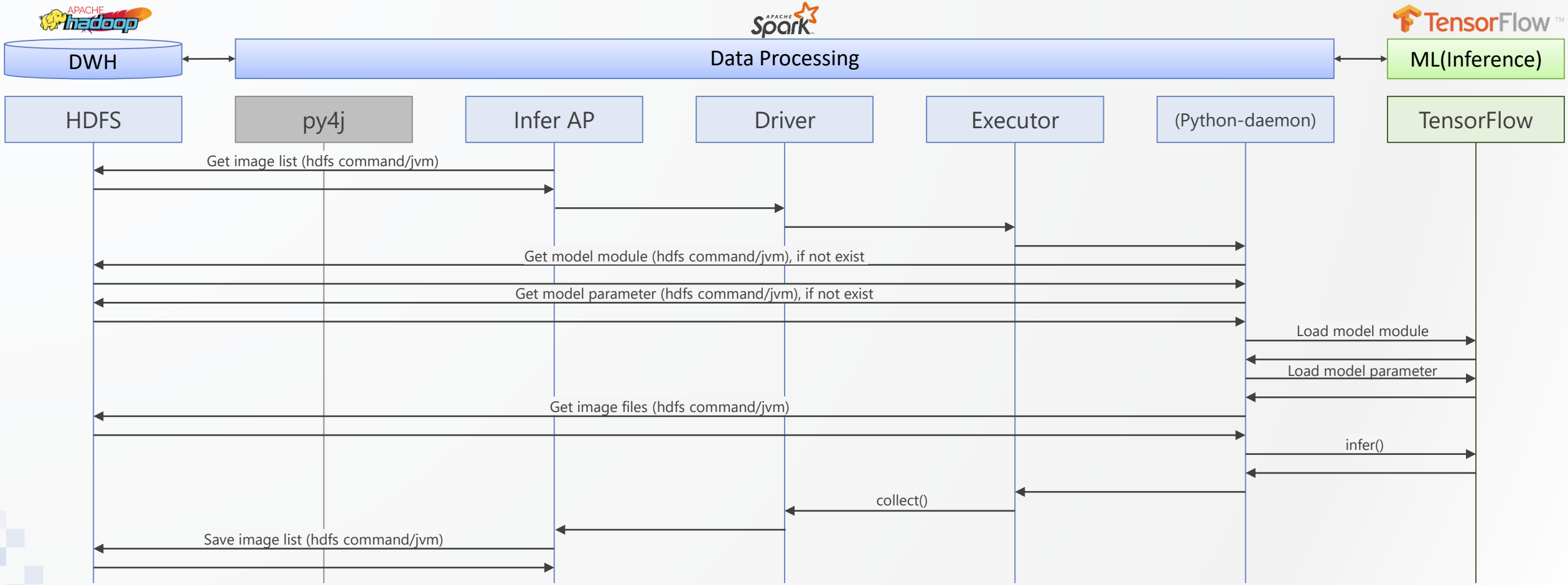


Figure: Sequence Diagram in case of **TensorFlow API-based** inference processing (25->22)

© 2019 NTT DATA Corporation

LL: Integration of Apache Spark and TF

The complexity of processing logic changes depend on whether the inference processing implementation is Spark API based or the TensorFlow API based.

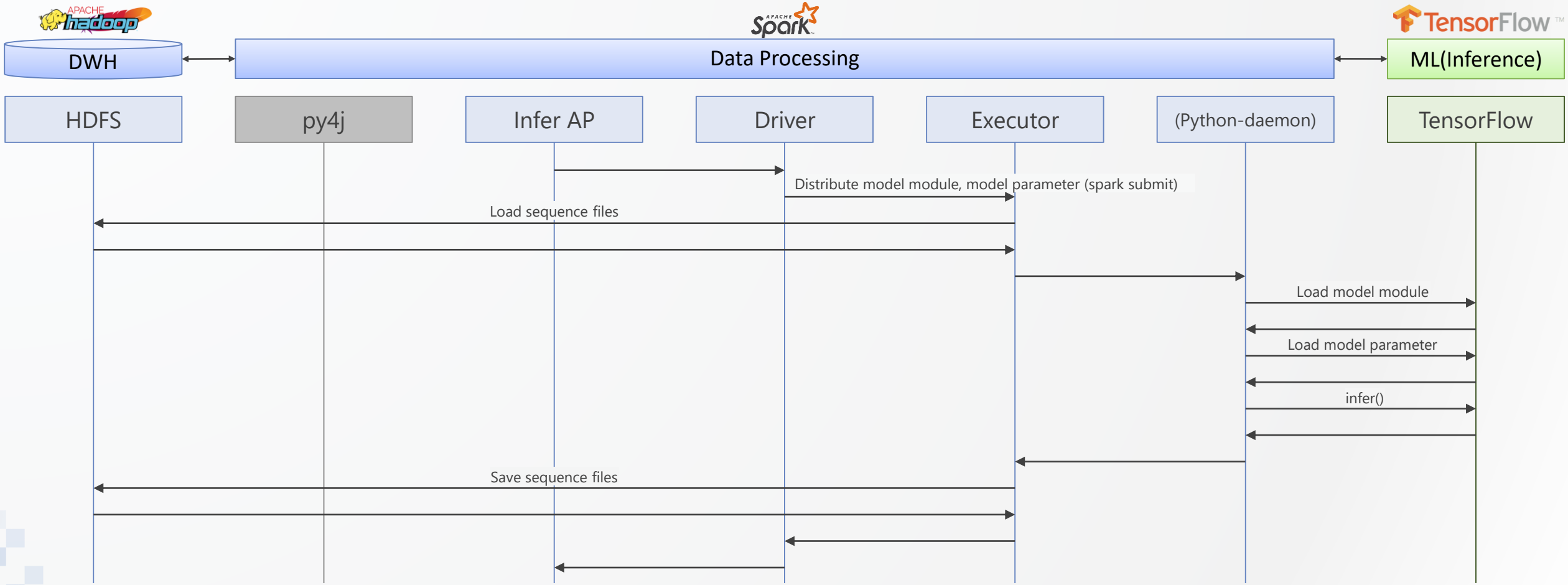


Figure: Sequence Diagram in case of **Apache Spark API-based** inference processing (25->22->16)

© 2019 NTT DATA Corporation

Summary



- **We introduced the challenges, approach and solutions in our ML project - Giant Hogweed Eradication.**

- **We shared our lesson learned to adopt ML in SYS through the project.**

  ➡ Future Work: Try & Error and get a better practice of other items in Scaled ML Systems.

© 2019 NTT DATA Corporation



© 2019 NTT DATA Corporation