

セミナー情報

SEMINARS

外部主催セミナー

2019年

▶ 「Strata Data Conference New York 2019」

■開催概要

会期 : 2019年9月26日
会場 : Javits Center(New York)
主催 : Cloudera, O'REILLY

■セッション詳細

Agenda一覧

■NTTデータセッション 講演内容

【4日目】「Deep Learning Technologies for Giant Hogweed Eradication」

時 間 4:35pm-5:15pm

NTTデータ OSSプロフェッショナルサービス 土橋 昌, 梅森 直人

Giant Hogweed is a highly toxic plant originating in the Western Caucasus. It has spread across Central and Western Europe and there are sightings of Giant Hogweed reported from North America, too. Landowners are obliged to eradicate it, due to its toxicity and invasive nature in Europe. However, it is difficult for landowners to find and remove Giant Hogweed across large areas of land since it is a very cumbersome manual process.

To automate the process of detecting the Giant Hogweed by exploiting technologies like drones and image recognition/detection using Machine Learning is an effective way to address this problem.

In this presentation, we show you how we designed the architecture, how we took advantage

of both of Big Data and Machine / Deep Learning technologies and lessons learned through this project. For example, we integrated a drone, Apache Hadoop, Apache Spark and TensorFlow to achieve the usability, flexibility and scalability for both of data engineers and data analysts. We talk about why this integration was needed for us, technical challenges from the view point of enterprises and tips to leverage the above open source software.

▶ 「Data Engineering Meetup #1」

■開催概要

会期 ： 2019年6月12日(水)

会場 ： Social Coding Studio of NTT Software Innovation Center(グランパークタワー)

主催 ： Databricks

■NTTデータセッション 講演内容

「Spark Meetup Tokyo #1」

時 間 18:40-19:35

NTTデータ OSSプロフェッショナルサービス 猿田 浩輔, 都築 正宜, 田中 正浩

▶ 「USENIX OpML'19 登壇・参加報告会」

■開催概要

会期 ： 2019年6月6日(木)

会場 ： 豊洲センタービル 10F INFORIUM (NTTデータ)

主催 ： 株式会社NTTデータ

■NTTデータセッション 講演内容

「A Distributed Machine Learning For Giant Hogweed Eradication」

時 間 18:00-19:30

NTTデータ OSSプロフェッショナルサービス 梅森 直人

▶ 「USENIX OpML'19」

■開催概要

会期 ： 2019年5月20日(月) 11:10～

会場 ： Santa Clara, CA, USA

■NTTデータセッション 講演内容

「A Distributed Machine Learning For Giant Hogweed Eradication」

時 間 11:10am-11:30am

NTTデータ OSSプロフェッショナルサービス 土橋 昌, 梅森 直人

▶ 「Data Engineering Meetup #1」

■開催概要

会期 : 2019年3月19日(火) 19:00～

会場 : LINE株式会社オフィス(JR新宿ミライナタワー)

会場 : LINE株式会社

■NTTデータセッション 講演内容

「Spark 2.4 & 3.0 - What's next -」

NTTデータ OSSプロフェッショナルサービス 猿田 浩輔

▶ 「Hadoop / Spark Conference Japan 2019」

■開催概要

会期 : 2019年3月14日(木)

会場 : 品川区立総合区民会館 きゅりあん

会場 : 日本Hadoopユーザー会

■NTTデータセッション 講演内容

Keynote 10:00-12:00

「The upcoming Spark 3.0: What's Next」

Databricks Xiao Li 氏(Spark PMC member), NTTデータ 猿田 浩輔 (Apache Sparkコミッタ)

会場B 16:05-16:40

「Apache Kafkaって本当に大丈夫? ～実際にいじめてみたのでお伝えします～」

NTTデータ 土橋 昌

2018年

▶ 「Strata Data Conference in New York」

■開催概要

会期 : 2018年9月11日～13日

会場 : Javits Center(New York)

会場 : Cloudera, O'REILLY

■NTTデータセッション 講演内容

【3日目】「Best Practices to Develop an Enterprise Datahub to Collect and Analyze 1TB/day Data from a Lot of Services with Apache Kafka and Google Cloud Platform in Production」

時 間 4:20pm-5:00pm

リクルートライフスタイル 林田 賢治氏

NTTデータ OSSプロフェッショナルサービス 佐々木 徹

Recruit Group is one of the largest web service providers in Japan. It has a lot of services covering diverse business fields such as travel and restaurant reservation, human resource services, POS systems, etc. Analyzing application logs collected from these various services enables us to provide more insightful services for individual/corporate customers. Our rough estimations showed the log size to be around 1TB per day and the number of servers/instances to collect logs from to be 1000+ in the future. We had to design a platform that can handle all these ever changing requirements. Therefore, we started a project to collect and analyze all the application logs generated by these services efficiently and easily. As the first step of this project, we developed a platform that handles extensive logs from upstream applications and transfer to downstream ones in an efficient and effective manner. This platform is based on the “Datahub” architecture and utilizes Apache Kafka in central for high performance and scalability. The Kafka cluster is developed on Google Compute Engine and some managed services in Google Cloud Platform such as Google BigQuery, Pub/Sub, etc, are also utilized for analysis. We faced with quite a few number of technical problems during developing this platform and successfully solved each one of them with smart solutions. In this session, we will introduce some of the critical problems that anyone could face when developing a similar platform and lessons, know-hows, and best practices we learned from this experience as below.

- How to collect application logs from a lot of services easily
- How to manage schema evolution of each log and adapt new schema to each analysis platform
- A reference network architecture for datahub to connect from a lot of existing services