How to simplify data scientist's day in huge company

IBM, Igor Khapov
Khapov.Igor@ru.ibm.com
Data scientist's workflow

1. **Know use case**
2. **Find data**
3. **Persist data**
4. **Extract features**
5. **Train model**
6. **Run model**
7. **Deploy & monitor**

- **Application**
- **Predictions**
- **Data Steward**
- **Dev. & prototype**
- **Production**
- **Lead Data Scientist**
- **Data Science Engineer**

Experiments → Train model → Extract features → Persist data → Know data → Find data → Run model → Deploy & monitor → Application
Create and use services

Bluemix (Cloud Foundry)

Apache Spark as a Service

Deploy automation

Symphony

- Resource manager
- Tenant manager
- Workload manager
- iPython + Spark compute nodes

SoftLayer Cloud Object Storage

OpenPOWER

CloudFoundry

CloudNativeCon China 2018
Create integrated solutions

**Authoring Tools**

- Jupyter
- R Studio
- SPSS Modeler
- Decision Optimization
- Cognos Dashboards
- Data Refinery
- Model builder
- Watson API tools

**Machine Learning Runtimes**

- Apache Spark
- dmlc XGBoost
- Spark MLlib

**Deep Learning Runtimes**

- TensorFlow
- torch
- Caffe
- Keras

**Scalable & modern infrastructure**

- Docker
- Kubernetes
- PowerAI
- NVIDIA
Use k8s to create dynamic infrastructure

Kubernetes cluster
- Admin dashboards
- Monitoring
- Pod spawner
- k8s master components

Control nodes

NGINX
- DSX UI
- Models

Compute nodes

Redis
- Gluster
- Object Storage
- Cloudant

Storage nodes

Data stores
- LDAP
- Spark cluster
- Object Storage
- Data stores (DB2, HDFS …)

Client environment

Kubernetes cluster

Data Science Experience

Services
- Analytics libraries
- Jupyter
- Spark

Client environment

KubeCon
CloudNativeCon China 2018
How to simplify data scientist's day

- Use existed and create your own ‘as a Service’ models
- Create end to end integrated solutions
- Provide expensive hardware resources as a service
- Use container orchestrators to create dynamic infrastructure
- Experiment with new technologies
- Keep calm and **have fun**