DSX Local

Overview & Roadmap

July 2018

DSXL Labs

https://ec2-54-88-58-154.compute-1.amazonaws.com

user01 - user20 / IBMdsxl!

Value of a Data Science Platform

Explore at scale

- Scale out on-demand
- No Dev-ops/engineering setup

Reproducibility

- Process of tracking
- Reproduce results easily

Secure

- Governed Access
- Administration capabilities

Collaborate

- Understand what's been done
- Share and accelerate learning

Publish Efforts

- Models as APIs out of the box
- Avoid Engineering re-work

Discovery to Production

- Minimal efforts
- Seamless scale
- Integration with business process

Open

- Use desired tool of choice
- Interoperability across tools

Review Results

- Stakeholder review
- Via Dashboards/Static reports

Monitoring

- QA/QC on-demand
- Retrain

IBM DSX - *the* Experience for Hybrid Data Science

Develop & Collaborate

Notebook servers and RStudio for interactivity & data visualizing with Python, R, & Scala for coder data scientists

Deploy at Scale

Spark parallelizes & accelerates data science tasks.
Automate, Deploy scoring servers at scale and monitor model health

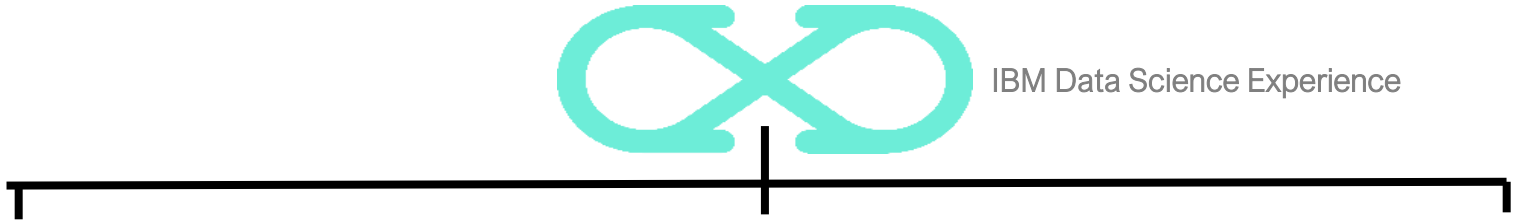
Watson Studio on IBM **Public Cloud**	DSX **Desktop**	DSX **Local** on Private Cloud

▪ PayGo consumption with as-a-service delivery, up & running in seconds
▪ Integrated with IBM Spark-as-a-Service & Watson Machine Learning as a service for compute,
  ▪ IBM Object Store & other cloud services for data,
▪ Publish and collaborate in the cloud
	▪ Easily installed on your laptop or PC
	  ▪ Won't scale beyond the hardware available on your machine
	▪ Access to RStudio, Zeppelin and Jupyter notebooks, and one small Spark worker operating locally on your machine
	▪ Load CSV data files into Data Frames
		▪ Scalable DSX cluster deployed on your private infrastructure
		  ▪ Dockerized containers via Kubernetes & deployable on IBM Cloud Private
		▪ DSX Local can also deploy with Hortonworks Data Platform & IIAS appliance on-premises
		▪ LDAP for user management and authentication, easy collaboration with Projects enabled by git.

➢ Use DSX *wherever* it makes sense for you (or where your data is) - and you can easily collaborate on the *same* project across these environments

and with v1.2: SPSS Modeler for Clickers and Visual Coders

Introducing IBM Data Science Experience Local



Community

- Find tutorials and datasets
- Connect with other Data Scientists
- IBM ML Hub for expert assistance
- Open Source evangelism
- Fork and share projects, samples

Open Source

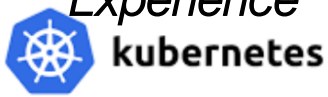
- Code in Scala/Python/R/SQL
- Zeppelin & Jupyter Notebooks
- RStudio IDE
- Anaconda distribution
- Add your favorite libraries

IBM Added Value

- Projects and Version Control , for dev->test->prod continuous engineering
- Relational & Hadoop data sources connectivity
- Machine Learning & Deep Learning - manage/monitor & deploy models
- Spark-in-DSX and Remote Spark (Hadoop) as well as Python & R based Analytics, ML.
- Publish notebooks and other assets, Host R Shiny Apps, schedule jobs
- Compute Elasticity support, manage CPU/GPU & memory resources
- Data Science Elite team

DSX is an Open Platform

Data Science Experience



- ✓ **Add your favorite libraries**
- ✓ **Publish Open APIs for secure ML applications**

Get started easily..

Learn

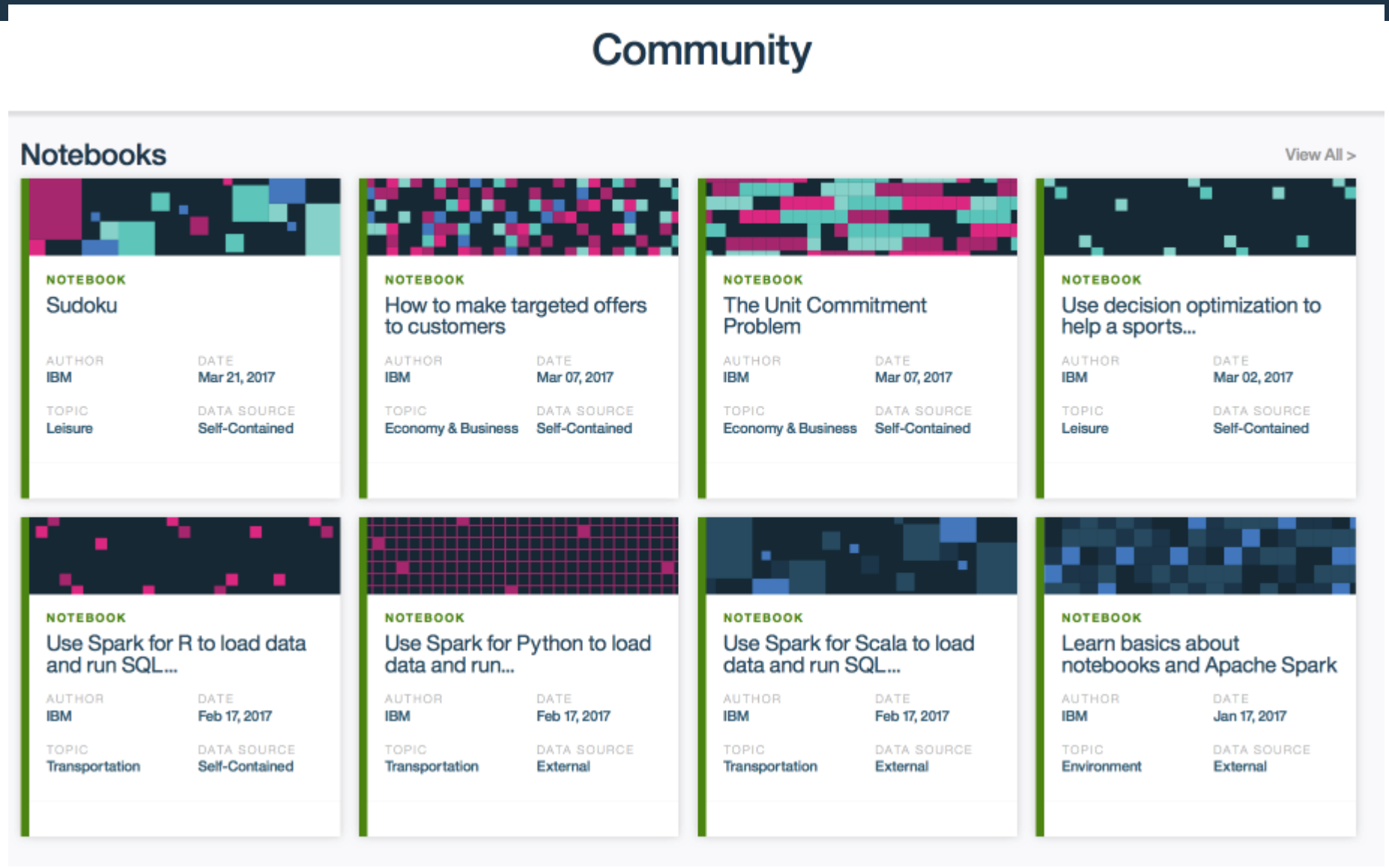
Connect to Enterprise data sources easily

Collaborate

Working on cluster safer than desktops for leader

Safe behind the firewall

Big SQL, Db2 (warehouse/z/LUW),
Hive for HDP, HDFS for HDP
Hive for Cloudera (CDH)
HDFS for Cloudera (CDH)
Informix, Netezza, Oracle



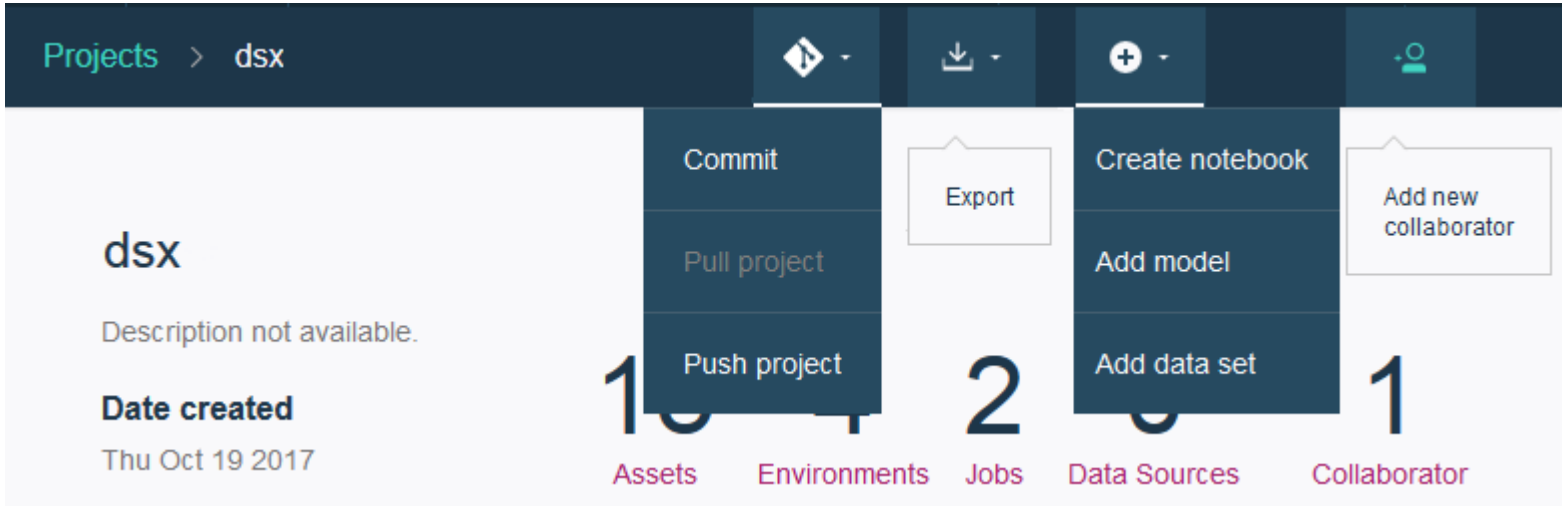
and explore

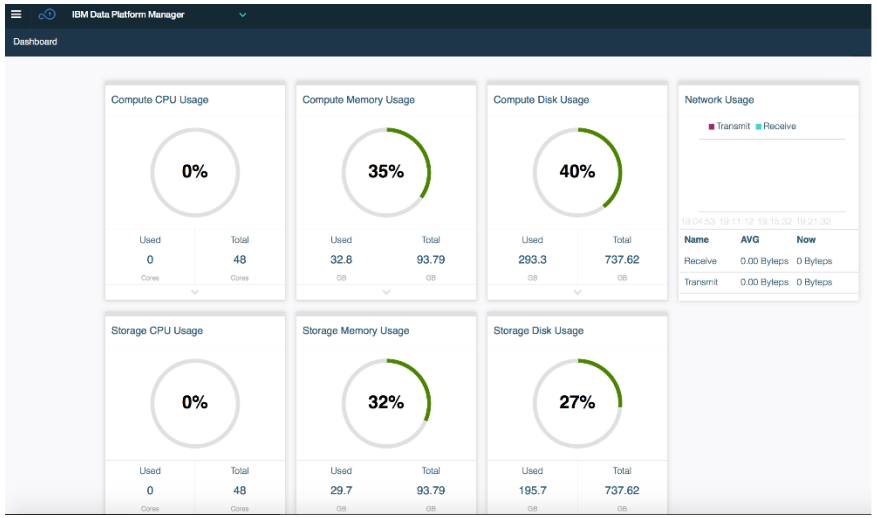
DSX Local simplifies distribution of team work based on skills

DSX Local increases knowledge sharing and knowledge retention

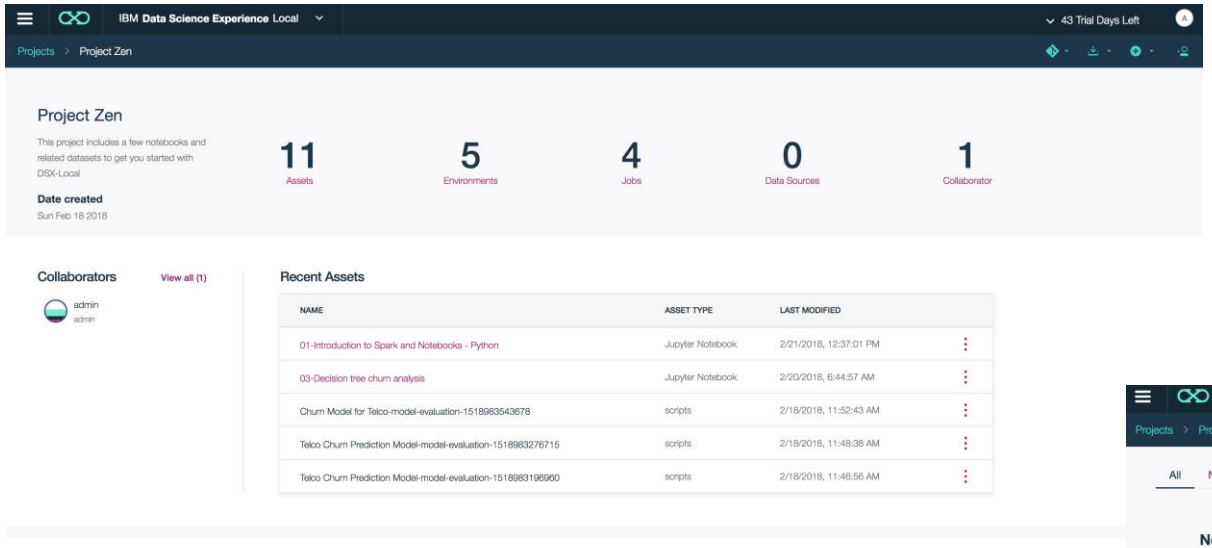
Currently based on open source notebooks, productivity tools in the future

DSX Local simplifies cluster management for teams

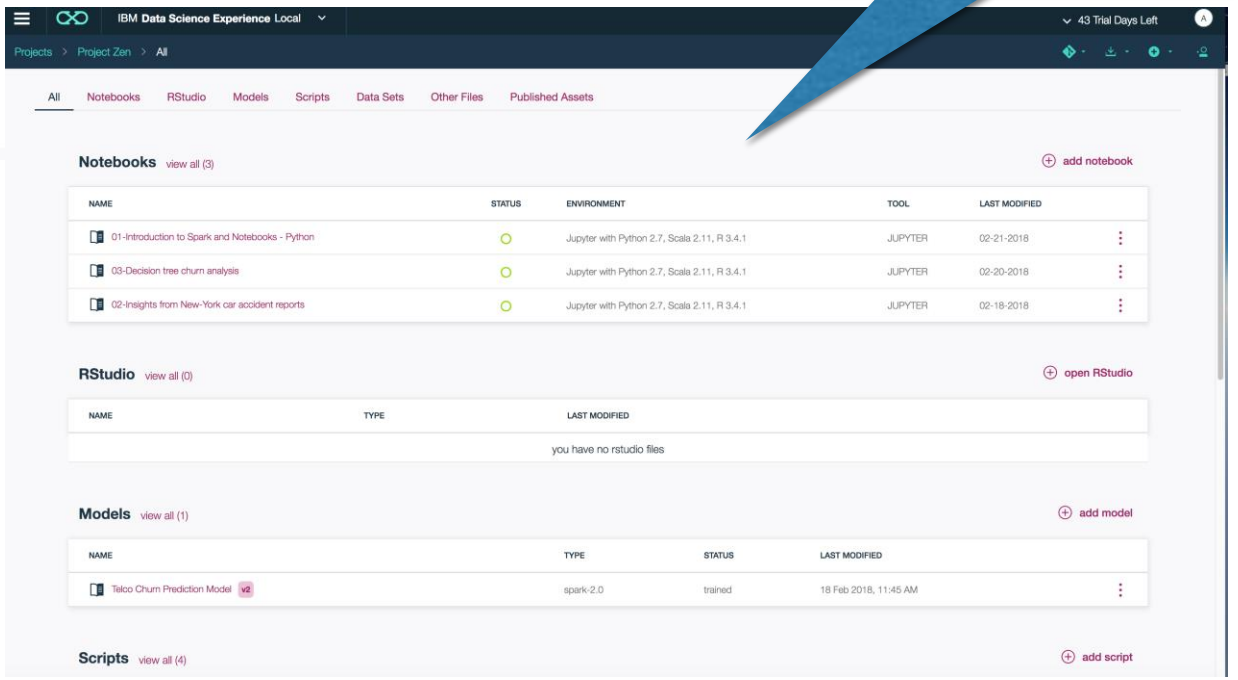




Data Science is a team sport



Collaborate within git-backed projects



Build & Collaborate

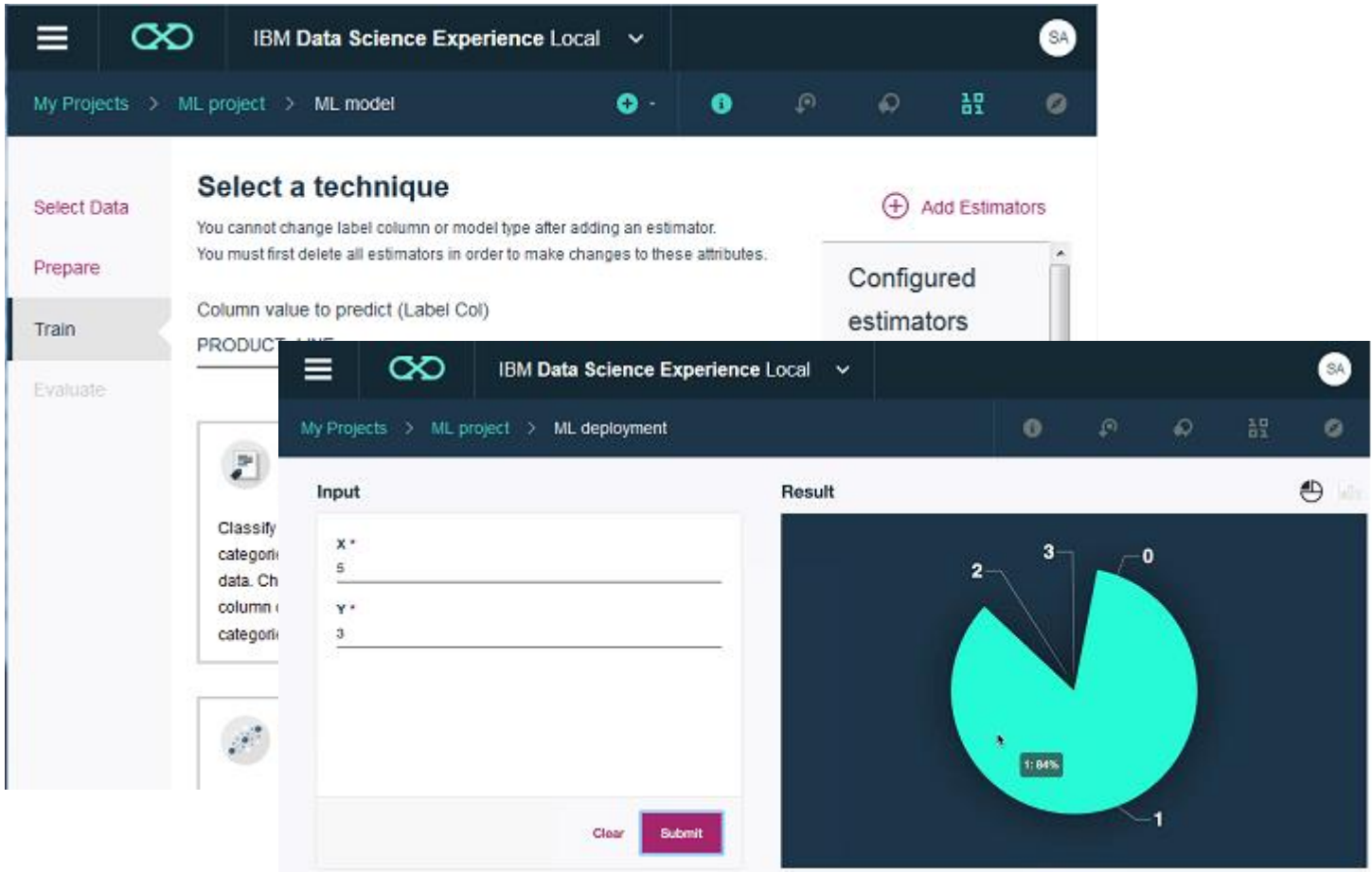
Facilitate creation of machine learning models

Facilitate deployment of models as API end-points

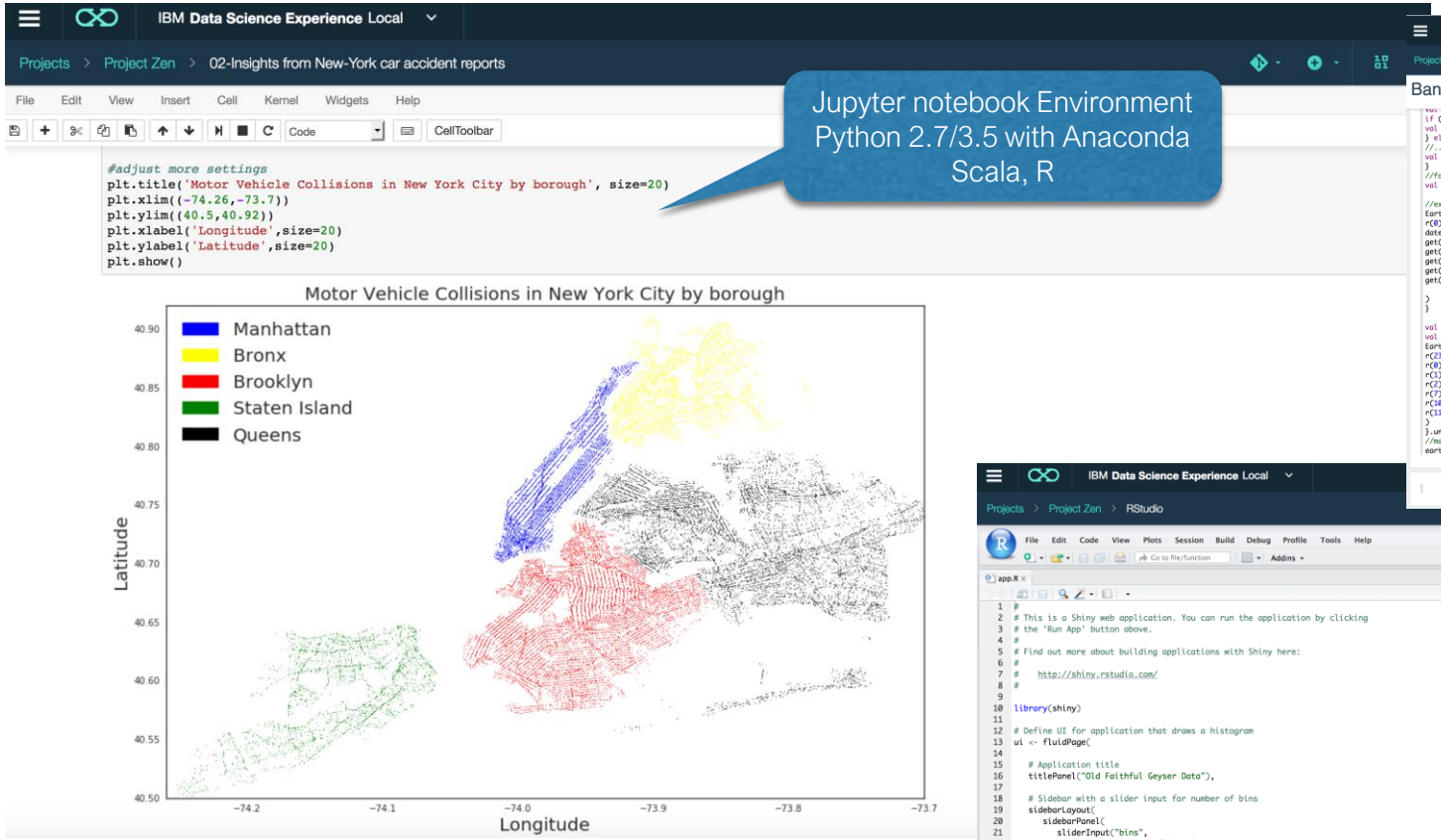
Automation of Batch Scoring, Training and Evaluation scripts as schedulable jobs

GIT integration to collaborate with engineers in their favorite environment

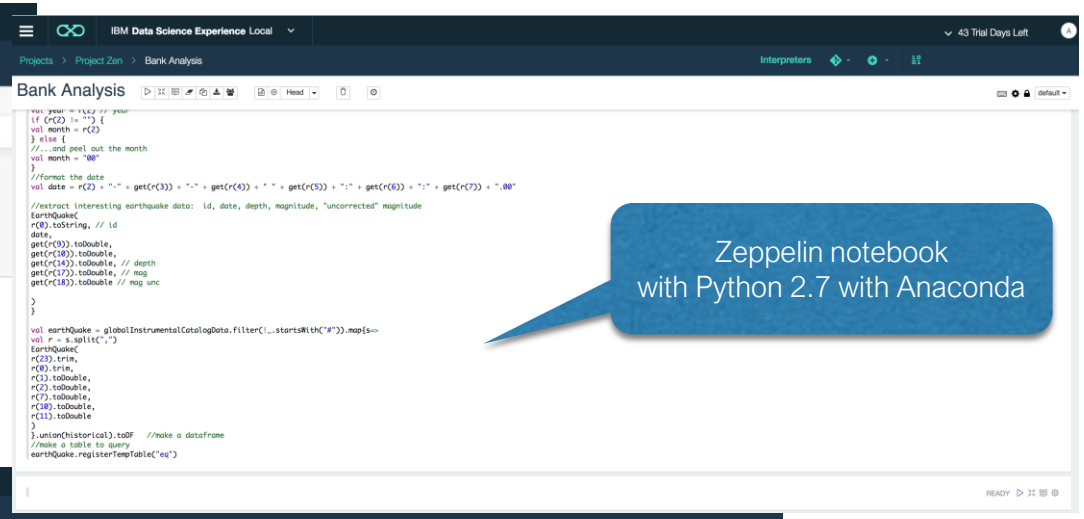
Publish content to others in pdf / html / R-Shiny app



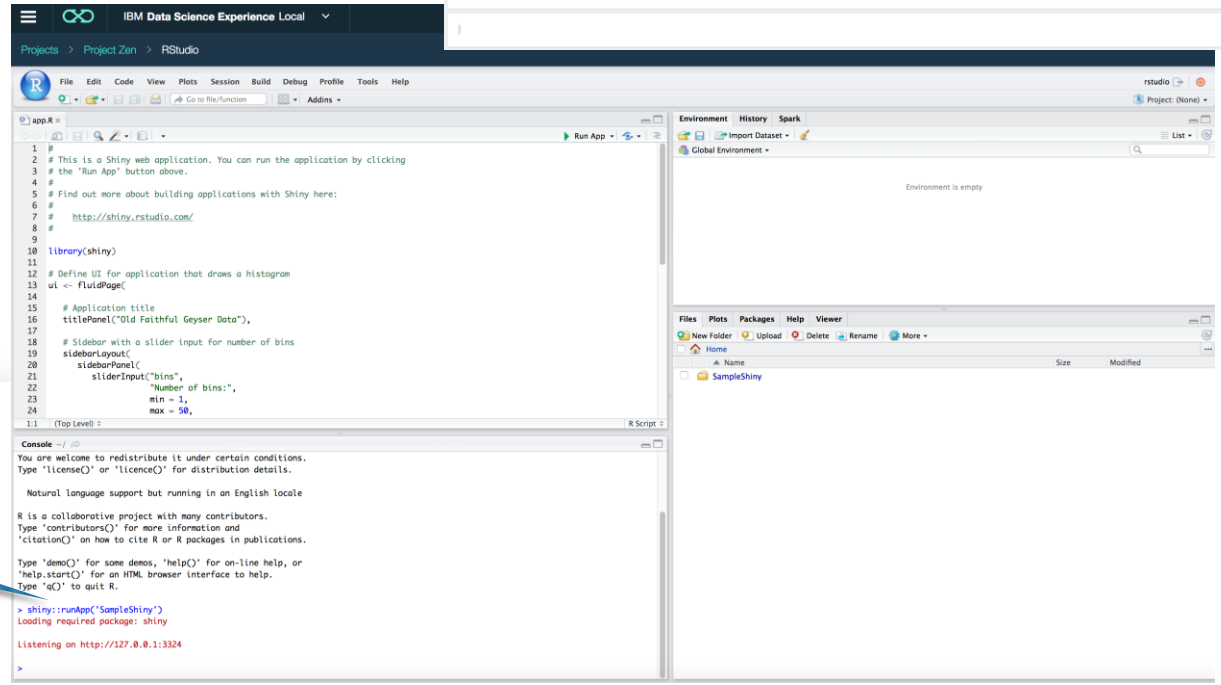
Essential tools for Data Scientists



Jupyter notebook Environment
Python 2.7/3.5 with Anaconda
Scala, R

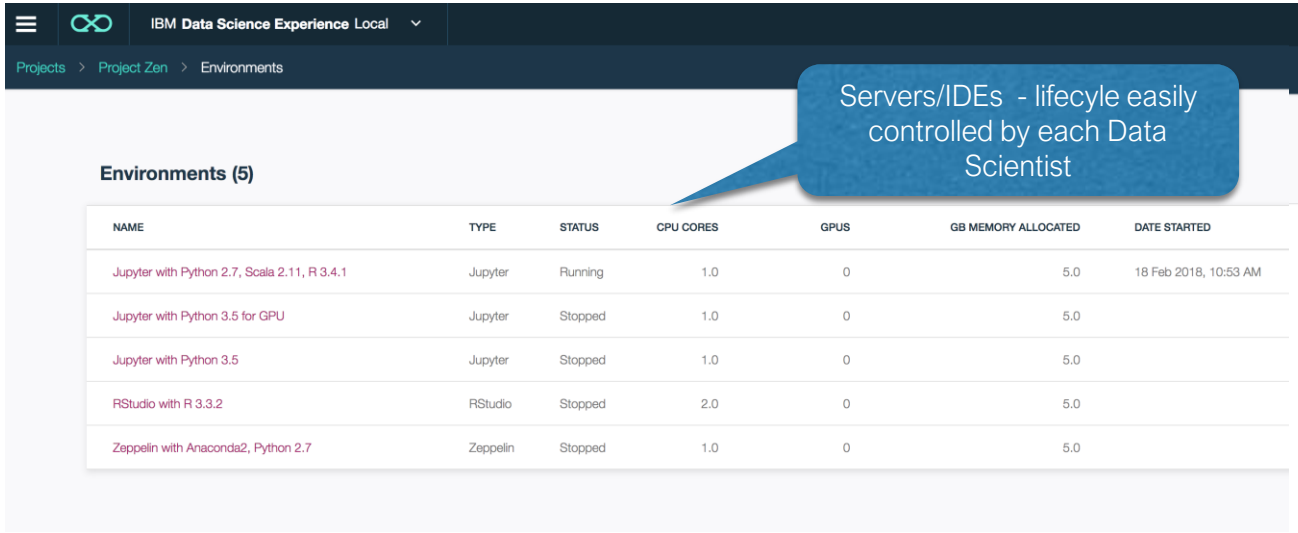


Zeppelin notebook
with Python 2.7 with Anaconda

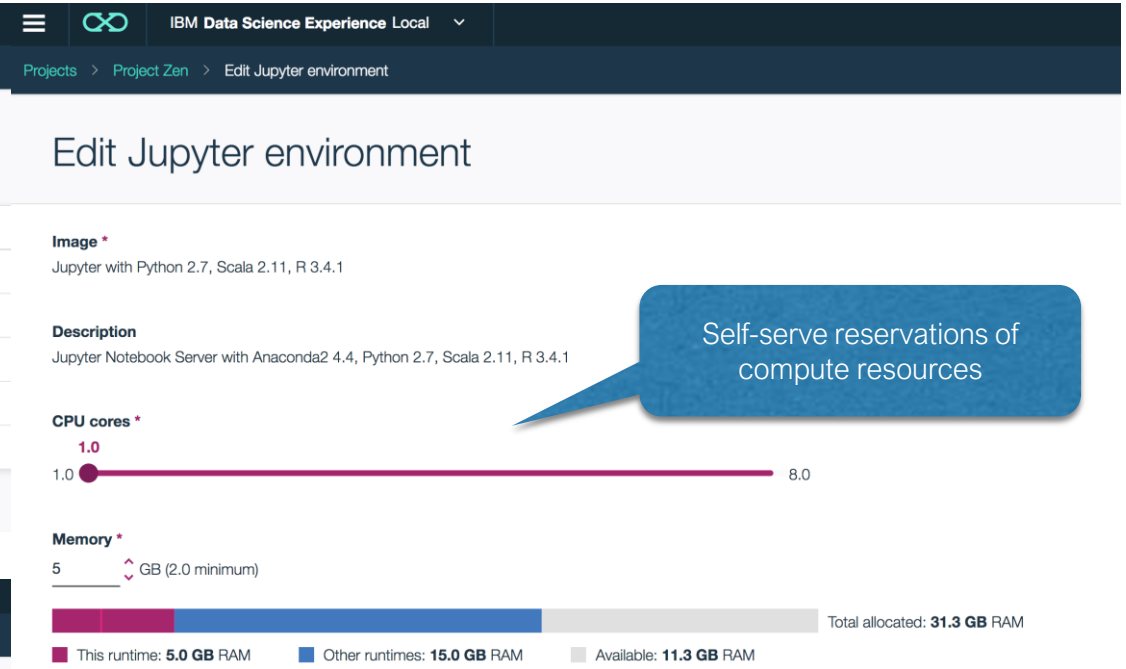


R Studio Environment with >
300 packages, R Markdown, R
Shiny

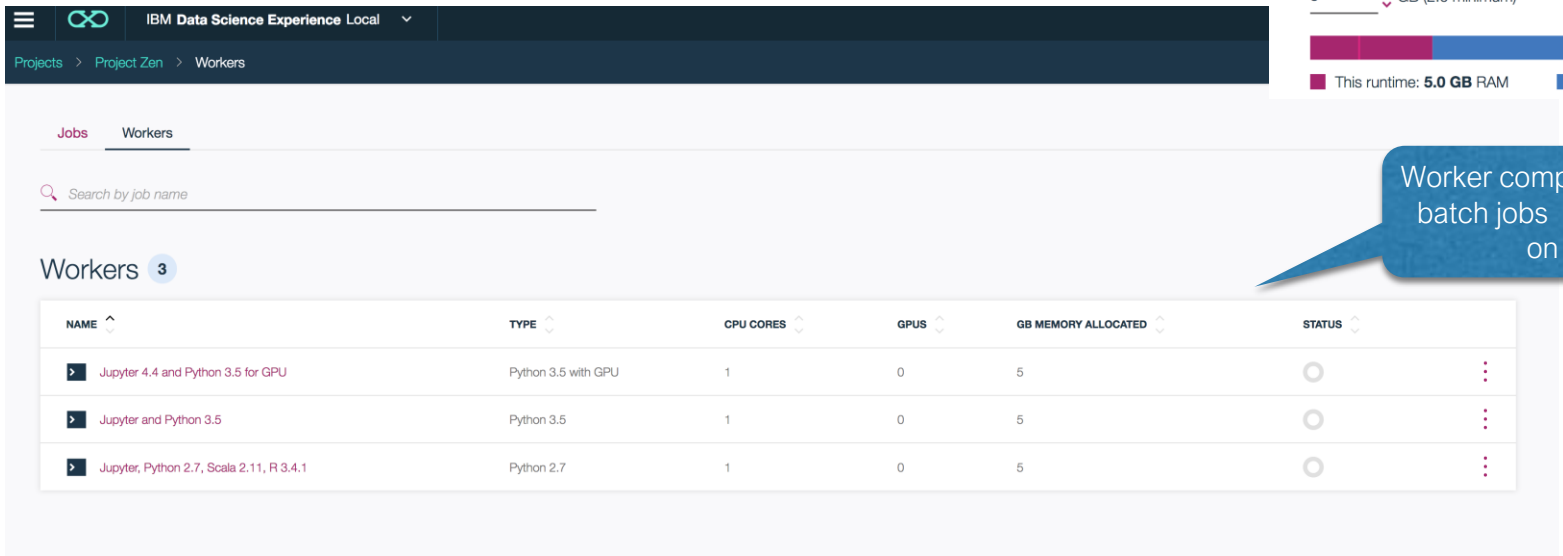
Self serve Compute Environments



Servers/IDEs - lifecyle easily controlled by each Data Scientist



Self-serve reservations of compute resources



Worker compute resources – for batch jobs run on-demand or on schedule

Environments are essentially Kubernetes pods – with High Availability & Compute scale-out baked in
(load-balancing/auto-scaling is being planned for a future spring)

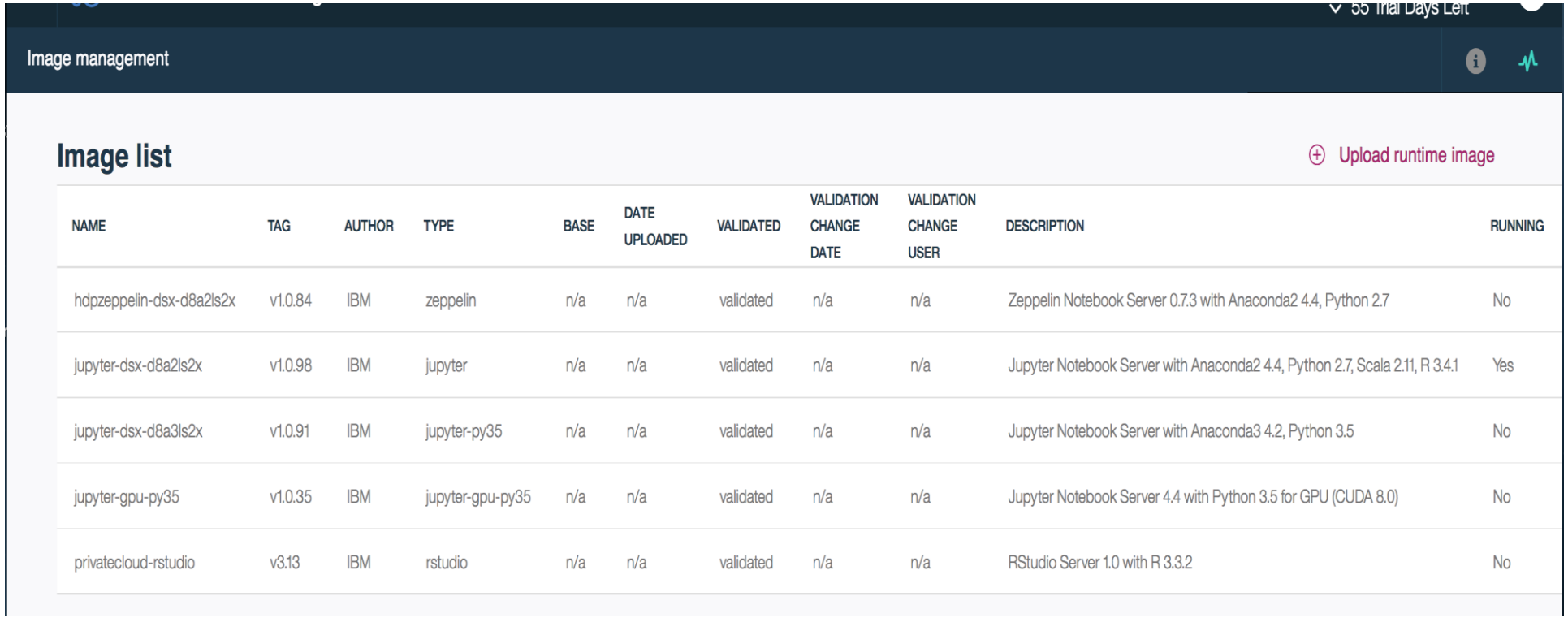
Extend ..
– Roll your own Environments

- Add libs/packages to the existing Jupyter, Rstudio , Zeppelin IDE Environments or introduce new Job “Worker” environments

https://content-dsxlocal.mybluemix.net/docs/content/local/images.html

DSX Local provides a Docker Registry (and replicated for HA) as well.

These images get managed by DSX and is used to help build out custom Environments



NAME	TAG	AUTHOR	TYPE	BASE	DATE UPLOADED	VALIDATED	VALIDATION CHANGE DATE	VALIDATION CHANGE USER	DESCRIPTION	RUNNING
hdpzeppelin-dsx-d8a2ls2x	v1.0.84	IBM	zeppelin	n/a	n/a	validated	n/a	n/a	Zeppelin Notebook Server 0.7.3 with Anaconda2 4.4, Python 2.7	No
jupyter-dsx-d8a2ls2x	v1.0.98	IBM	jupyter	n/a	n/a	validated	n/a	n/a	Jupyter Notebook Server with Anaconda2 4.4, Python 2.7, Scala 2.11, R 3.4.1	Yes
jupyter-dsx-d8a3ls2x	v1.0.91	IBM	jupyter-py35	n/a	n/a	validated	n/a	n/a	Jupyter Notebook Server with Anaconda3 4.2, Python 3.5	No
jupyter-gpu-py35	v1.0.35	IBM	jupyter-gpu-py35	n/a	n/a	validated	n/a	n/a	Jupyter Notebook Server 4.4 with Python 3.5 for GPU (CUDA 8.0)	No
privatecloud-rstudio	v3.13	IBM	rstudio	n/a	n/a	validated	n/a	n/a	RStudio Server 1.0 with R 3.3.2	No

Automate ..

IBM **Data Science Experience** Local

43 Trial Days Left

Projects > Project Zen > Jobs

Jobs | Workers

Search by job name

Jobs – trigger on-demand or by a schedule. such as for Model Evaluations, Batch scoring or even continuous (re-) training

Jobs 4

create job

NAME	TYPE	SOURCE ASSET	CREATOR	DATE CREATED	SCHEDULED TO RUN	LAST RUN	STATUS
Churn Model for Telco-1518983543678	Model evaluation	scripts/Churn Model for Telco-model-evaluation-1518983543678.py	admin	18 Feb 2018, 11:52 AM	On demand	18 Feb 2018, 11:52 AM	
Telco Churn Prediction Model-1518983276715	Model evaluation	scripts/Telco Churn Prediction Model-model-evaluation-1518983276715.py	admin	18 Feb 2018, 11:48 AM	On demand	18 Feb 2018, 11:48 AM	
Telco Churn Prediction Model-1518983196960	Model evaluation	scripts/Telco Churn Prediction Model-model-evaluation-1518983196960.py	admin	18 Feb 2018, 11:46 AM	On demand	18 Feb 2018, 11:46 AM	
Telco Churn Prediction Model-1518982994199	Model evaluation	scripts/Telco Churn Prediction Model-model-evaluation-1518982994199.py	admin	18 Feb 2018, 11:43 AM	On demand	18 Feb 2018, 11:43 AM	

DSXL Labs

https://ec2-54-88-58-154.compute-1.amazonaws.com

user01 - user20 / IBMdsxl!

https://github.com/jpatter/DSX_Local_Workshop_V12/

Deploy, monitor and manage

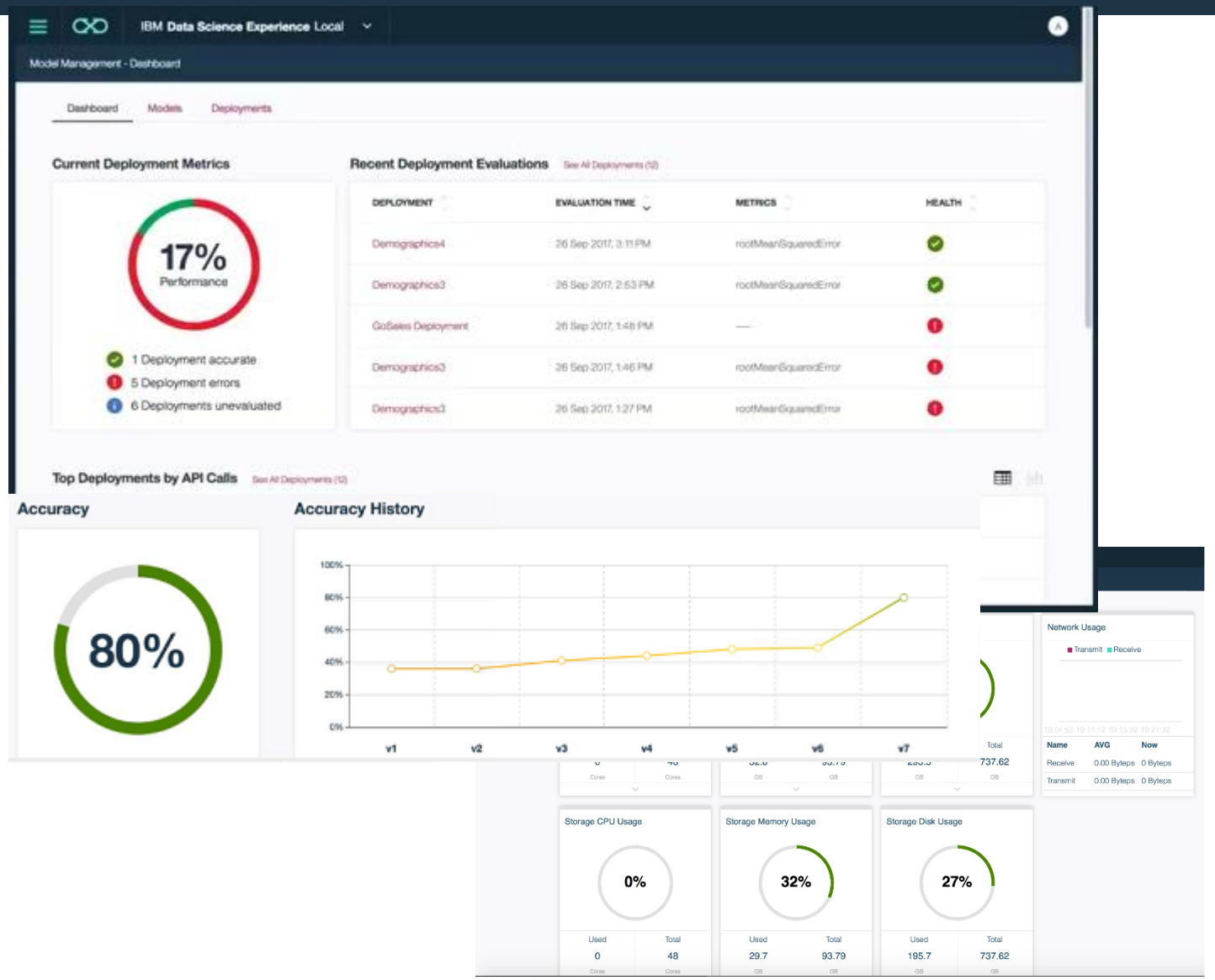
Monitor models through a dashboard

Model versioning, evaluation history

Publish versions of models, supporting dev/stage/production paradigm

Monitor scalability through cluster dashboard

Adapt scalability by redistributing compute/memory/disk resources

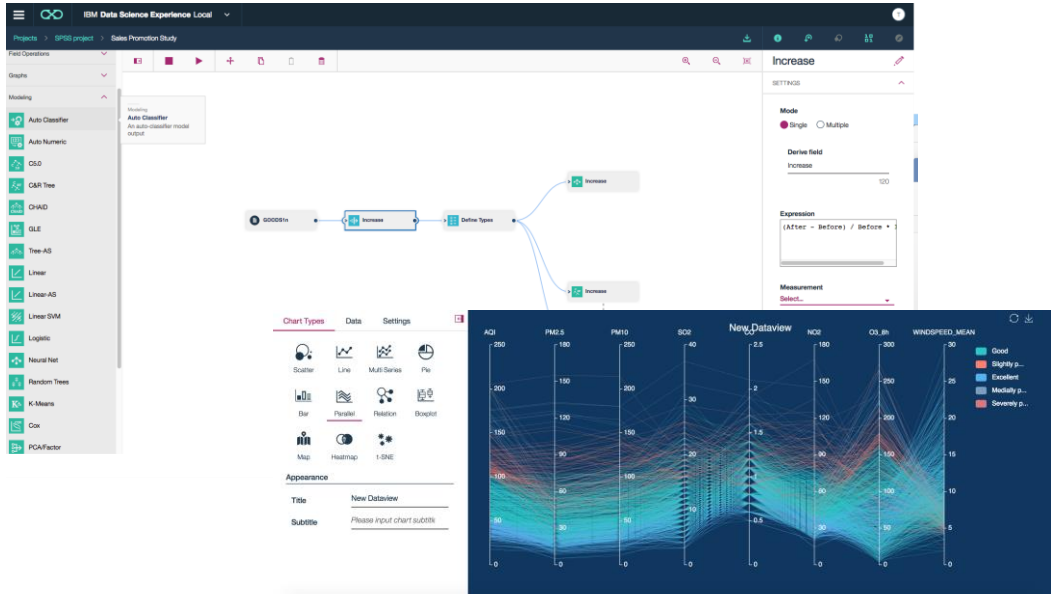


What's new in v1.2.0

- "MMD" – Deployment Manager
- Hadoop Integration Service
- SPSS Modeler (Officially GA) - as an add-on
- Decision Optimization community edition pre-packaged – full version as an add-on
- Data Refinery (Beta)
- Python & R Script Editor
- R Machine Learning models support
- Projects: Tagging, Commit History, BitBucket support, Enhanced Tree view
- Jobs Enhancement: Run SPSS Modeler Flows, Stop/Cancel running jobs
- Custom JDBC driver: to connect to generic JDBC capable sources
- Improved administration experience: adding jdbc drivers, key & certificate tasks, setup Livy end-points, manage Hadoop Integration Service end-points

SPSS Modeler for DSX

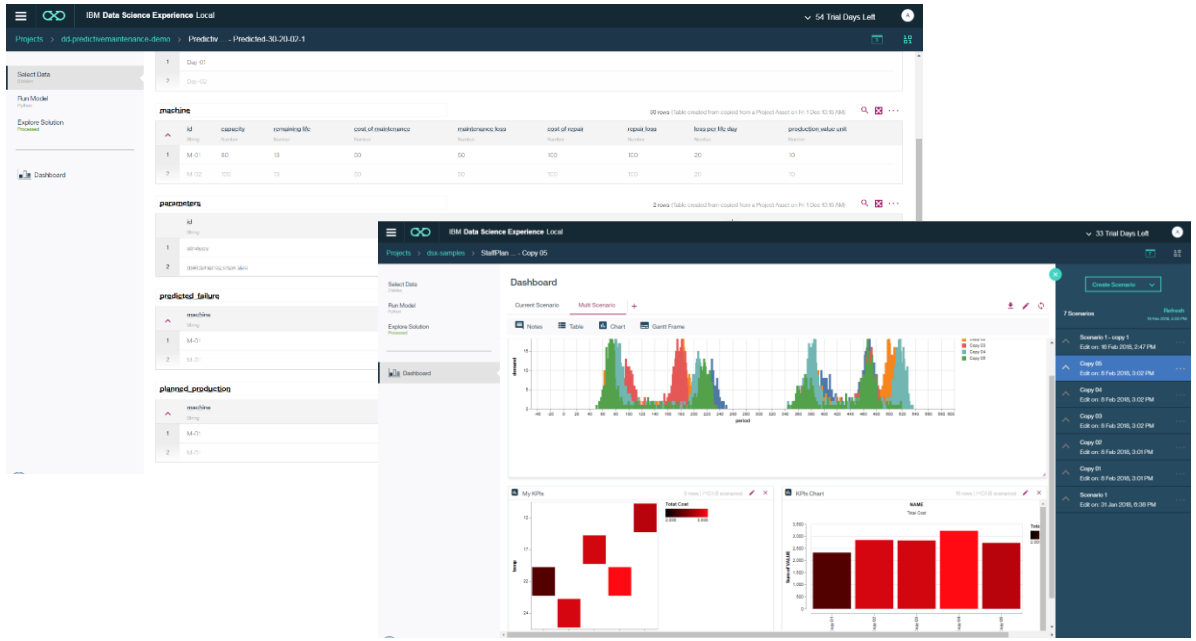
- **Pain Points:**
  - Lack of skills around coding
  - Environment for quick prototyping/experimentation
  - Easy path to deployment from visual productivity
- **Value Proposition:**
  - Visual productivity tool around data science
  - Quicker time to value
  - Inclusion of full-fledged data preparation and many machine learning algorithms



- **Features:**
  - Newly rebuilt interface with improved navigation and ease of use
  - Totally new interactive visualization
  - Ability to deploy results in Model Management and Deployment

Decision Optimization for DSX

- Pain Points:
  - Operationalize Data Science and drive higher ROI
  - One platform for predictive and prescriptive tools
  - Enable team on prescriptive tools and approaches
  - Productivity of expensive resources
- Value Proposition:
  - Transform insights from ML into actions using DO
  - Add-on to DSX local with a mix of predictive and prescriptive in unified environment
  - Community cards, tutorials, etc.
  - Validate models using dashboards and execute what-if



- Features:
  - Model building workflow
  - What-if analysis and dashboards
  - Beta: Modeling assistant (limited domains)

MMD: Introducing Manage Deployments feature

- Supports Dev->Test->Staging->Production CICD paradigm
  - Access control & Separation of duties
  - Enables a clear path for promotion of assets to production, while enabling Development to continue on separately.

- Dev: (DSX Local Projects)
  - Build & test out assets
    - Notebooks
    - Python & R Scripts
      - Run as a Job in DSXL cluster or against the Hadoop Integration Service
      - REST based tests against User authored Scripts run as a Web Service
        (invoke Python or R functions – for example for Custom "Scorers", to infuse custom data prep)
    - Models – Train & Test from scripts/notebooks
    - Shiny Apps – Build & Preview
    - SPSS Modeler Flows - Design & Test

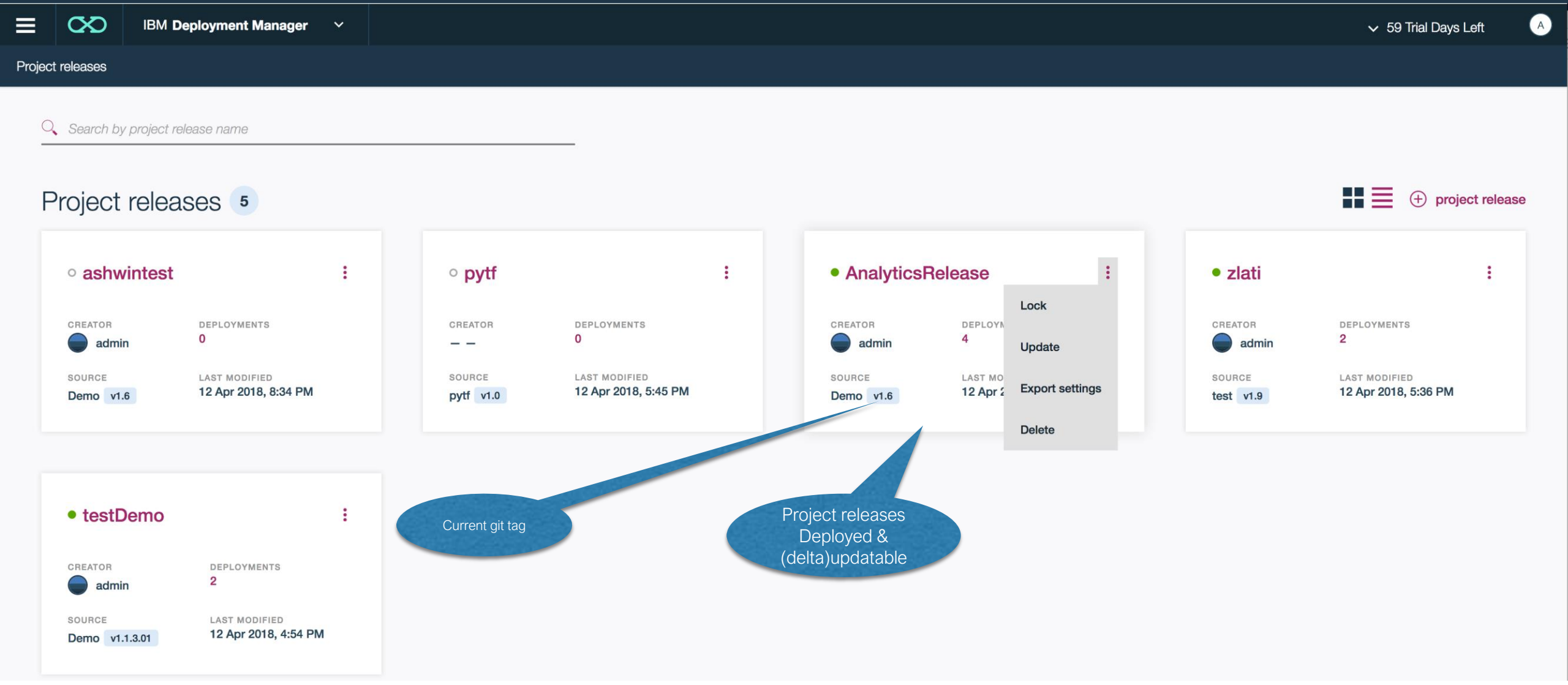
- A Project "Release" **Tag**
  - identifies a stable checkpoint in the history of the Project that marks its (possible) *promotion* to Production
  - implemented as a git tag (also applies to github/bitbucket repositories or exported project .zip/.tar.gz archives)

MMD: Introducing Manage Deployments feature

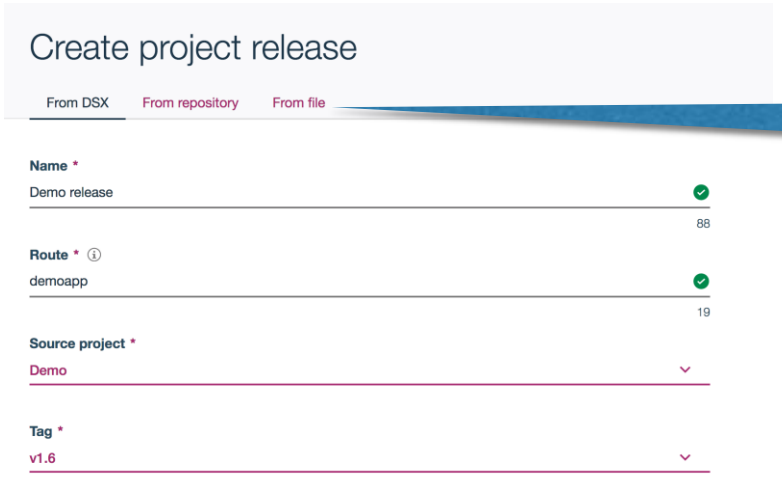
- Production with “MMD”: either in the same cluster as Dev or use a separate cluster
  - Pull from the release git tag(or export/import)
  - Create deployments
    - Create online scoring end-points for Models (or versions of a model)
    - Schedule Batch Jobs such as SPSS Modeler Flows, Python/R scoring or evaluations, against Hadoop etc.
    - Schedule execution & externalize Jupyter Notebooks
    - Externalize Shiny Apps
    - Expose User authored Scripts as a Web Service
      - API based access to Python or R functions –such as Custom Scorers with custom data prep
  - Define resource requirements & SLAs for Deployments
    - Per deployment – choose to “reserve” compute (or acquire when needed)
    - Load balancing and latency - choose to run replicas of Scoring servers, Scripts-as-Web Services for high throughput needs

- Variations:
  - Export a “Deployment Manifest” from a “Staging” MMD and Import - modify in a Production MMD
  - Use the same MMD instance for both staging and production – just create two “Project Releases” with different URL routes.

Deployment manager - Project Releases

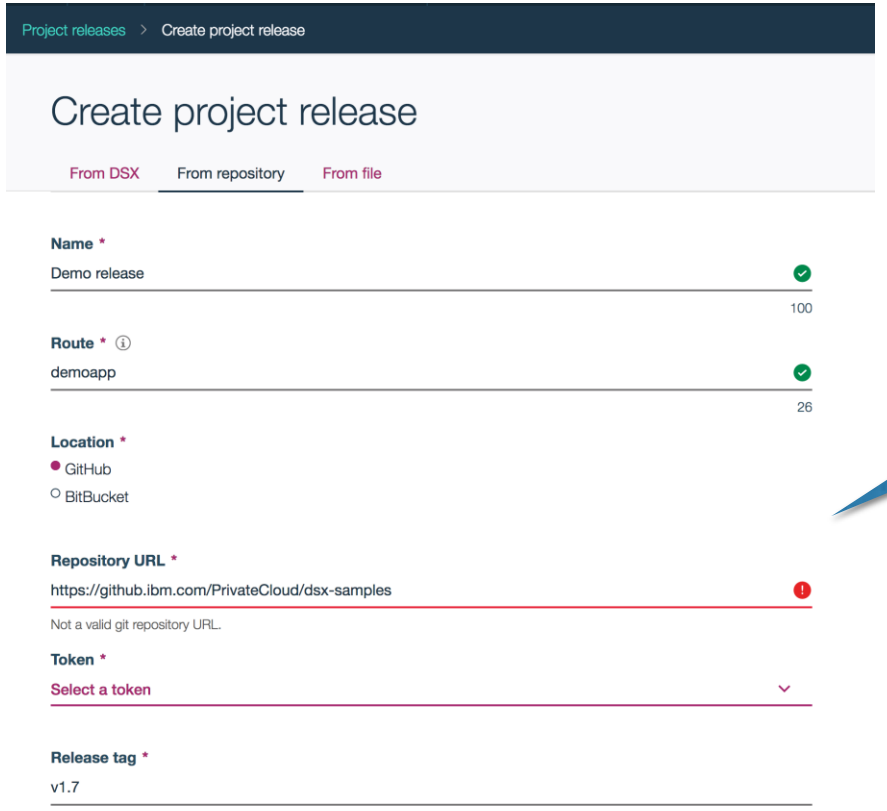


Bring in a new “release” to production



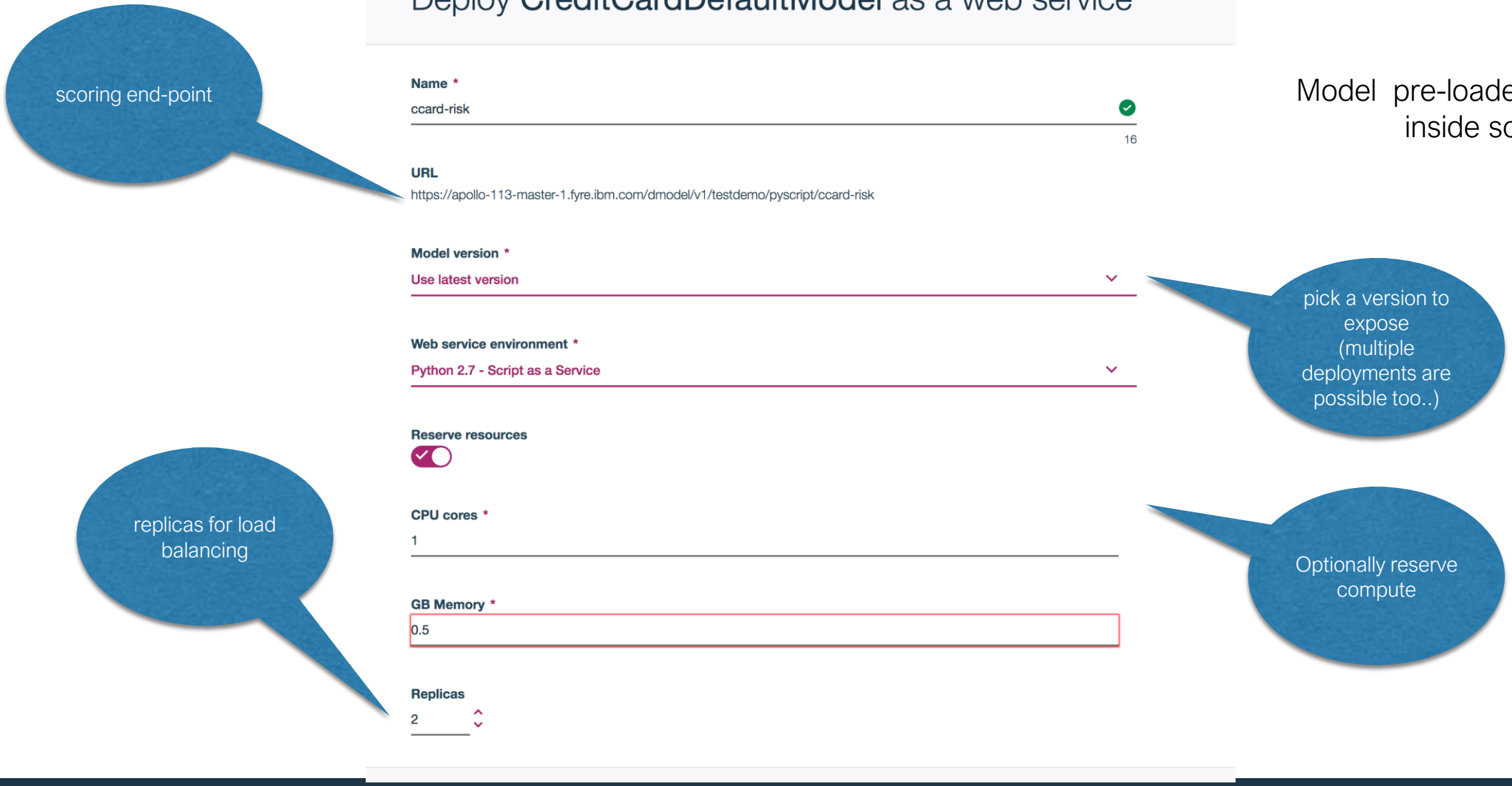
New Releases
- from a “Source” Project created from a .tar.gz package

New Releases
- from a “Source” Project in the same cluster



New Releases
- from a “Source” Project pulled from github/bitbucket

Expose a ML model via a REST API



scoring end-point

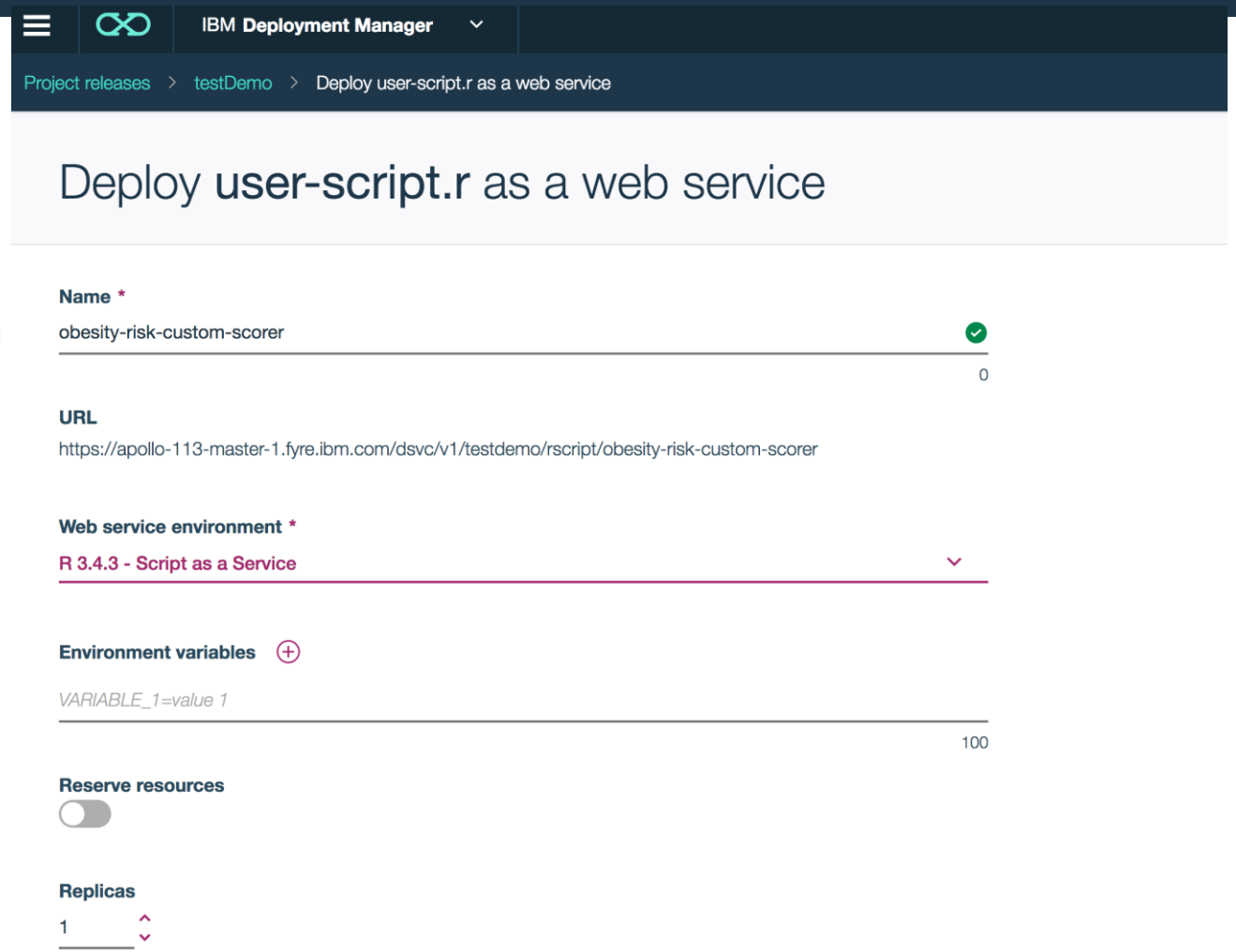
replicas for load balancing

pick a version to expose (multiple deployments are possible too..)

Optionally reserve compute

Model pre-loaded into memory inside scoring containers

Expose Python and R scripts as a Web Service



Custom scripts can be externalized as a REST service - say for custom prediction functions

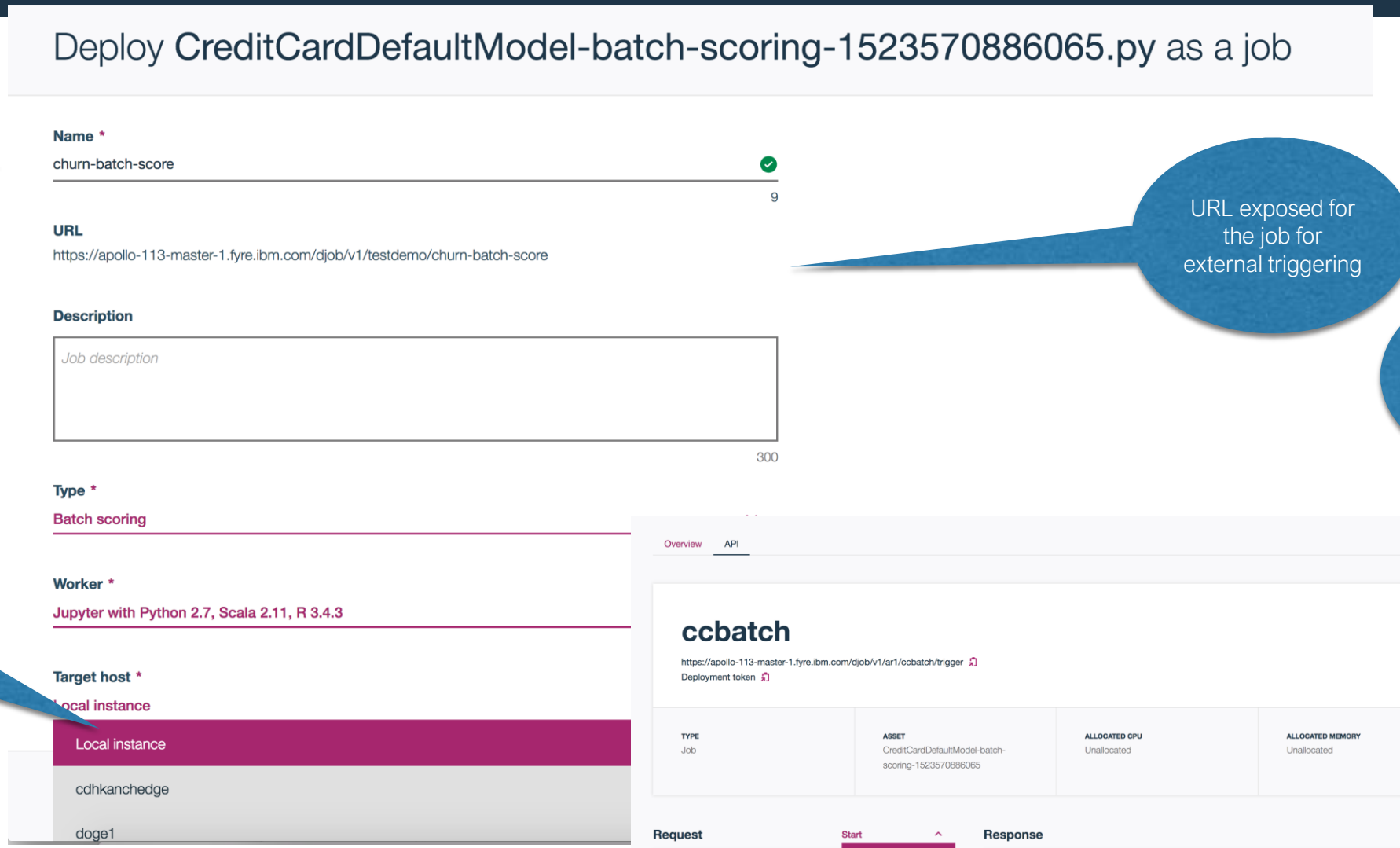
Deploy a script as a schedulable Job

Scripts can be deployed as a job..

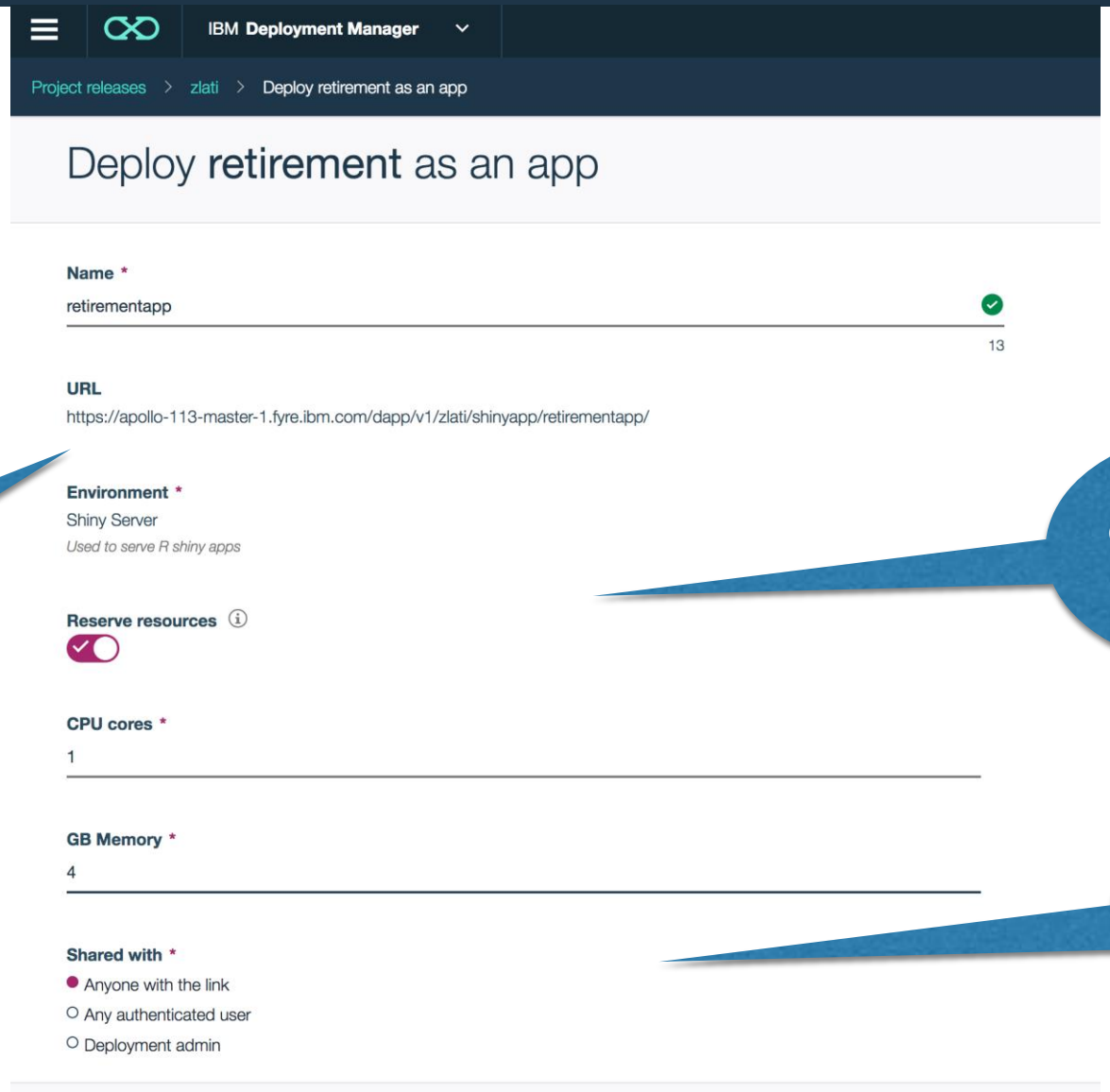
URL exposed for the job for external triggering

And use the convenient API panel to understand how to invoke the job..

Can be run against a DSX Hadoop Integration Service environment



Deploy a Notebook or R Shiny App

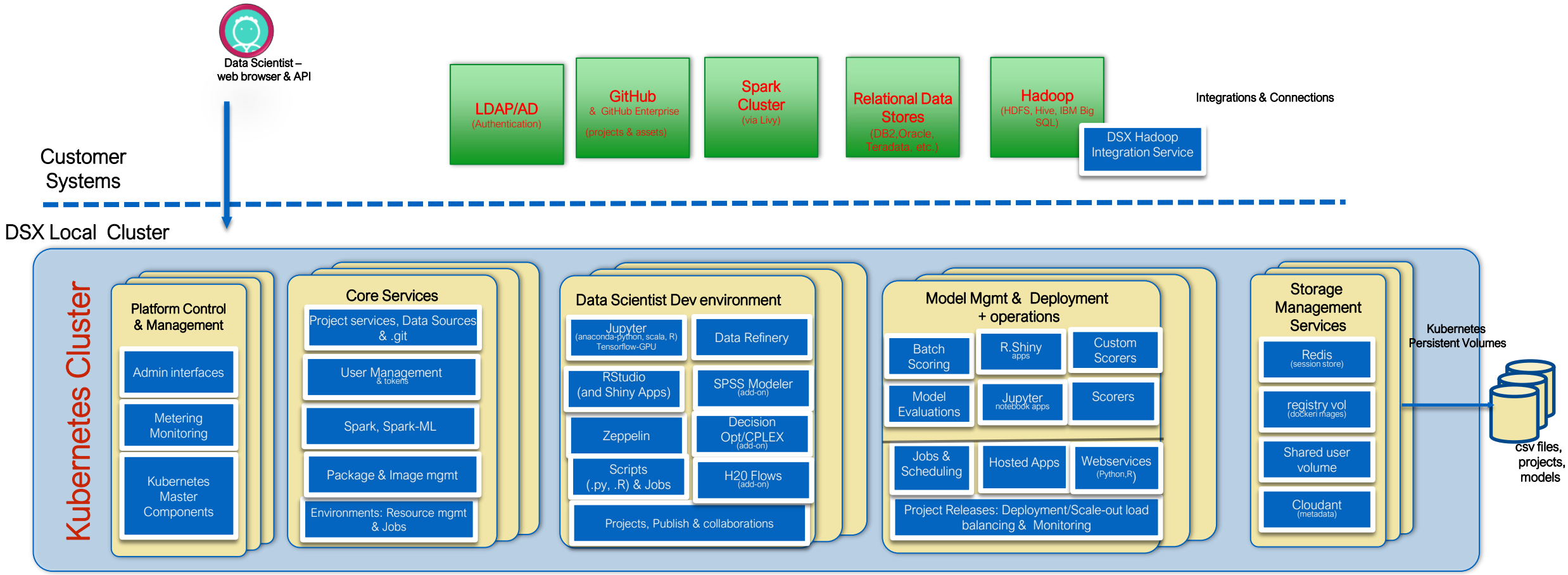


Host a notebook or a Shiny App

Optionally reserve compute

Define visibility rules

DSX Local Architecture overview

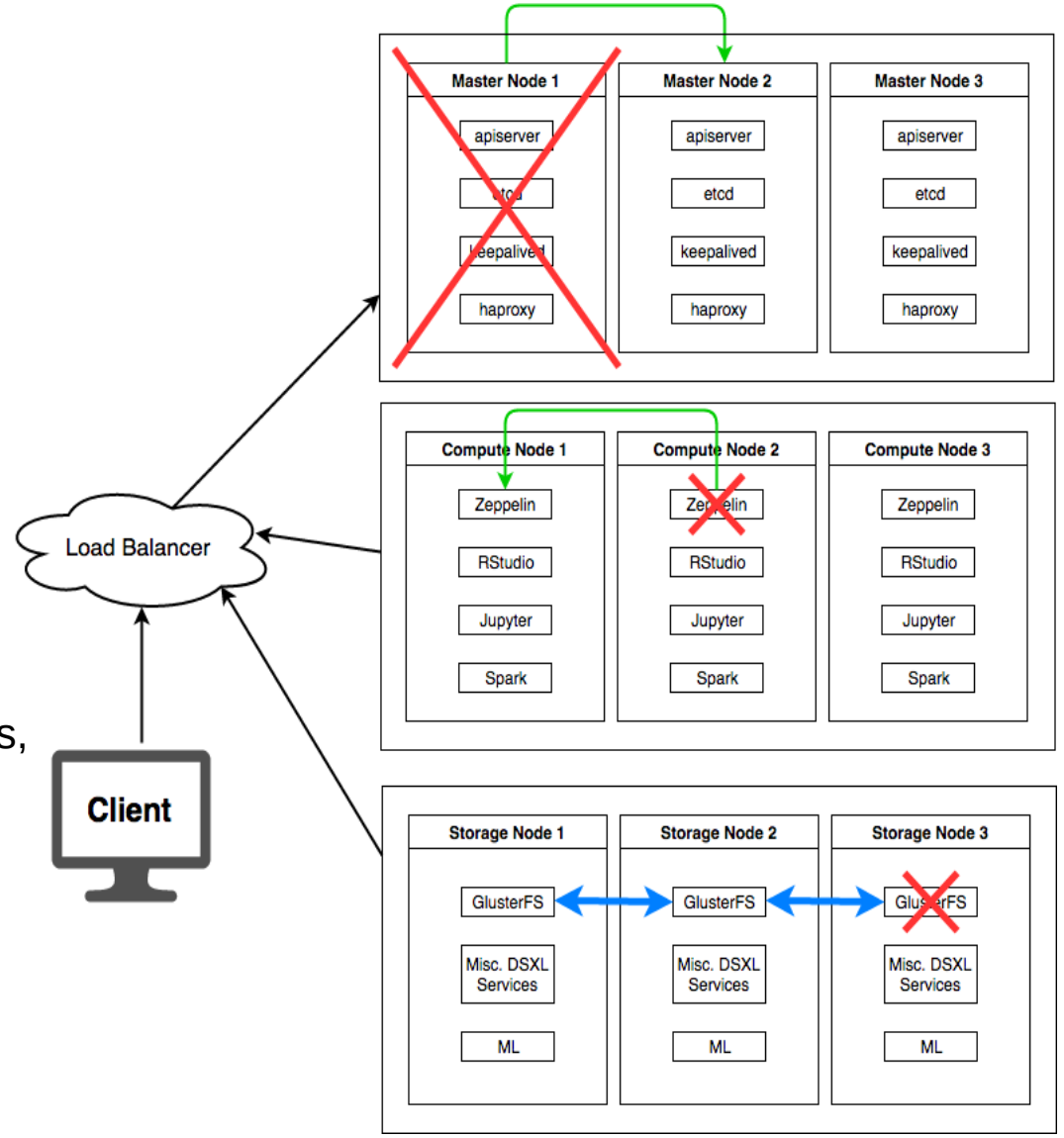


- Start with 3 nodes with HA enabled
- Expand as needed

Kubernetes cluster spread across multiple servers

HA in DSX

- **Platform HA**
  - By default, the cluster is configured for HA: minimum 3-node
  - Can tolerate failure of
    - 1 node in a 3-node configuration
    - 4 nodes in a 9-node configuration
- **Service HA**
  - For services that are deployed as pods, Kubernetes will monitor and redeploy
  - No session failover, which means that some services may be down for a few minutes



Roadmap Highlights

Strategy Highlights

Jan 2018 (Delivered)

New data connections to ease data access

- Support for IBM BigSQL on HDP as a data source
- Support for Cloudera data sources

Enhanced ML capabilities

- Train, Deploy and Monitor Scikit-learn, xgboost, Tensorflow+Keras and spark-ml models
- Automation - Batch Scoring and Evaluation scripts as schedulable jobs
- Model versioning and evaluation history
- Publish models

New Deep Learning and GPU support

- Runtime Environments can now be associated with GPU(s)
- Train Tensorflow+Keras models or execute other jobs that need GPUs in these Environments.

Collaborate

- Publish and share the URL to any content (pdf, html, csv etc.) with other users

Mar 2018 (Delivered)

IBM SPSS Modeler

- Refreshed GUI
- High priority nodes

Decision Optimization for Data Science integration

Model Management and Deployment package

- Deploy/Manage model versions
- Deploy SPSS Modeler Streams + batch scoring
- Deploy Notebooks, Scripts & Shiny Apps
- Add support for R based Models
- Management access control

Hadoop

- Python, R job push-down to Yarn in a secure Hadoop cluster (in-addition to Livy-Spark)

Q2 2018

SPSS Modeler enhancements

- Additional nodes available
- Extensions and scripting available with new GUI

Model Management and Deployment enhancement

- SPSS Modeler Streams – real-time scoring
- A/B testing & experimentation
- Scale-out Deployments - replicas & load-balancing(Scoring servers & Shiny Apps etc.)

Apache Atlas, IGC and Apache Ranger integration

Add-on/Extensions

- pick from a “market-place”, mix-n-map Environments easily

2h 2018

WEX integration - Text analytics – including support for ML annotators, classifiers, and custom annotators

HDP/Yarn 3.1’ s Kubernetes native support provides for Data Plane and DSX running directly on HDP-Kubernetes

Enable partner and third-party tool integrations

Support DL frameworks/libraries: DL4J, Caffe2, CNTK with GPU support.

Key scenario: Governed Enterprise Data Science

Exec

Sets goals & measures results

Data Engineer

Collects
Builds data lakes and warehouses
Gets Data Ready for Analytics

Development & prototyping

Production

App. Developer

Infuses ML in apps & processes

Real-time Apps & business processes

Problem Statement or target opportunity

Finds Data (from an enterprise catalog)

Explores & Understands Data

Predictions

Data Scientist

Refine

Collects, Preps & Persists Data

Extracts features for ML

Trains Model

Deploys & monitors accuracy

Auto-retrain

Analyzes
Explores, Shapes data & trains models

PRODUCTION
- *Secure & Governed*
- *Monitor & Instrument*
- *High Throughput*
  -Load Balance & Scale
- *Reliable Deployments*
  - outage free upgrades

Features, Lineage/Relationships recorded in Governance Catalog

Experiments

CDO (Data Steward)

Organizes

- Data & Analytics Asset Enterprise Catalog
- Lineage
- ***Governance*** of Data & Models
- Audits & Policies

Goal: Enterprise IaaS for Data Scientists

- Efficient Compute Resource Management for large-scale Analytics, Machine Learning and Deep Learning workloads
  - Enable Data Scientists to *procure* resources from a shared compute "grid" for any kind of activity – from interactive notebooks & IDEs to training Jobs or scheduled scripts and Apps.
  - All compute manifested as Docker containers/Kubernetes pods

- HDP/Yarn as *the* Resource Manager
  - Enable *all* workloads, whether Map Reduce or Spark Jobs or DSX/ML activities to be uniformly handled by the HDP/Yarn scheduler.
  - Manage Queue Priorities, balancing of workloads and scale-out for the whole cluster providing best utilization of all resources.

➢ Yarn and Kubernetes - the best of both worlds !