ORACLE
for Research

TECH TALK:

Data Science for Researchers

Friday, February 26th, 2020
10:30 AM US EDT

Rajib Ghosh
Global Senior Solutions Architect
Oracle for Research

Mike Riley
Cloud Solutions architect
Oracle for Research

ORACLE
for Research

TECH TALK HOUSEKEEPING

- Today’s webinar is being recorded. We will share the link to the recording with you via email after the event. The recording will also be made available to the Oracle for Research community.

- We invite your comments and questions, both about the tech topic being discussed and about the series more generally. Questions may be submitted using the Q&A box on your screen or you may ask questions directly using your microphone. When not asking a question, please mute your microphone.

- Questions may be asked during the presentation and we will also have a Q & A time at the end of the presentation when you can ask questions directly and engage in discussion.

- At Oracle for Research, we believe that research and innovation happen best when a diverse and thoughtful community is free to engage in respectful, compassionate, and open dialog. To that end, when asking a question or providing feedback, we ask that all participants be respectful, collaborative, and constructive.

Agenda

Recap and Asks from researchers	❑ Q & A – Data science options for research ❑ What data science options does a researcher have from Oracle ❑ Research use-cases , tooling, flexibility and data support do I have? ❑ What is cost effective and what makes the most sense?
Data science	Topics – to be covered ❑ Common data science research areas and use-cases ❑ Data Science PaaS vs IaaS Service vs OFR Images ❑ Supported algorithms and example solutions ❑ Oracle Machine learning (OML) and Accelerated data science (ADS) Advanced Topics of interest – not covered ❑ Deployment, migration and scaling best practices ❑ GPU accelerated data science ❑ Data science containers ❑ Using data science workflows - options
Demos	❑ Data science platform – key features ❑ Data science – autonomous ❑ Data science image

Data science use-cases for research

Research use-cases

- **Life -sciences**
  - Evaluate importance of parameters in functional cell and tissue simulations
  - Making drug-toxicity and drug safety predictions using ML/DL models
  - DNN models (DNNBrain) using FMRI for cognitive neuroscience predictions
  - Machine learning in genomic sequencing
- **Geo-spatial imaging**
  - Remote sensing imagery using Spatial and GIS ML/DL algorithms
  - AI based spatial correlation for Urban street images
  - AI platform for crowd sourcing visual data
- **Agro-life and farm applications**
  - Geo-spatial modeling with orthomosaic images and sfm algorithms
  - Statistical data cleaning of farm sensor data, outlier analysis
- **Structural Engineering**
  - Machine learning based model buildout
- **Quantum chemistry**
  - Quantum chemistry calculations using Machine & Deep learning
- **Linguistic**
  - Use AI/ML techniques to understand content and summarize
- **Social and Media**
  - ML and NN models for racial bias and sentimental analysis

Technologies

Common data science frameworks/ libraries

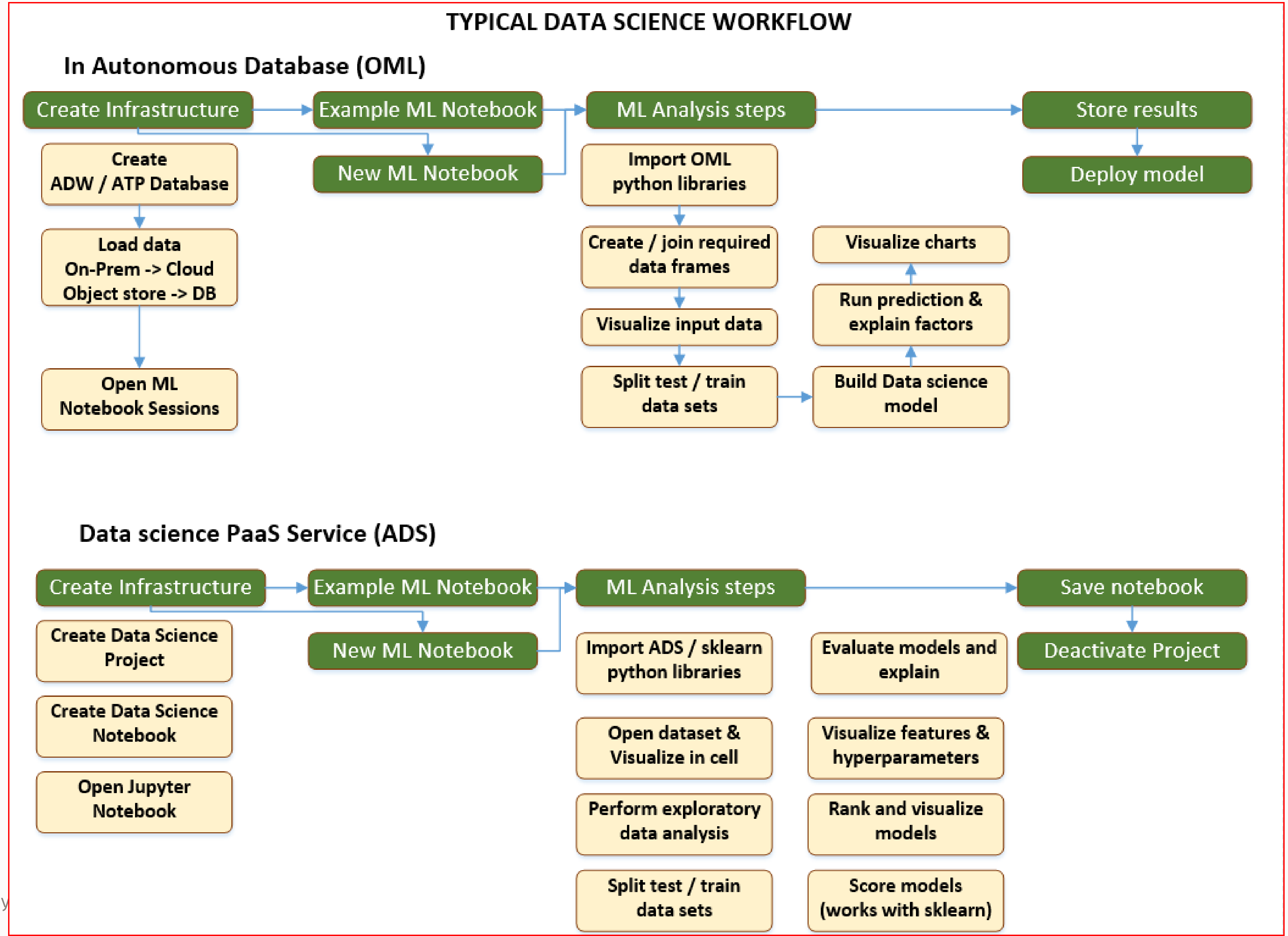
- TensorFlow
- Keras – Deep learning library
- Scikit-learn – Free python ML library
- Caffe / Caff2 – DL Framwrork
- Pytorch – Open source ML framework
- MXNet – Open source portable DL engine
- ONNYX - Transfer models across frameworks
- MATLAB – Visualizing DL / NN

Research areas

- Life sciences (Molecular dynamics)
  - TorchMD / HOOMD-TF / DPMD
- Brain Imaging – DNNBrain / Deeplabcut
- Geo-spatial – ARCGISPro ML/DL libraries
- Structural Engg – Viedt / maml / Yade
- Quantum Chemistry - Graph neural network
- Social Media – RNN / NLP sentiment analysis
- Text summarization models

Data science tools @ Oracle

All-in-one- data science Image	❖ Pre-built and tested data science image with popular tools and libraries installed ❖ Tested for NVDIA P100, V100 and A100 GPU architectures (both OCI VM and BM shapes) ❖ Support for open source JupyterLabs and Jupyter Notebooks ❖ OL7.x + CUDA 11.0.11 + cuDNN 7.6.5 as the base compatible version install ❖ Tools – Tensorflow 2, Pytorch, Keras, Mxnet, Scikit-learn, Seaborn, Pandas, Numpy, Matplotlib, pytorch
Data science platform	❖ Managed Data Science Platform as a service ❖ Self service and server less access to infrastructure for data science workloads ❖ End-to-end data science lifecycle workflow support for predictive models ❖ Data profiling/prep + Feature engg + Model Training + AutoML + Model explainability + production deployment ❖ Python libraries in conda environments that can be customized ❖ Multiple conda environment support with JupyterLab ❖ CLI and REST API automation and integration capabilities ❖ Accelerated data science library ❖ Project based and collaborative sharing if required ❖ Scalable across GPU shapes from test to high volume workloads ❖ Inactivate/activate projects to conserve resources keeping data intact to migrate/test across shapes
Oracle for Research Data science Image	❖ Conda based data science images with only basic conda installed ❖ Available on CPU and GPU Tensorflow versions on Ubuntu and OL7 ❖ Researcher can perform conda update to install desired CUDA versions & pytorch/mxnet ❖ Researcher can configure easily with any research specific deep learning libraries



TYPICAL DATA SCIENCE WORKFLOW

In Autonomous Database (OML)

Create Infrastructure

Create ADW / ATP Database

Load data On-Prem -> Cloud Object store -> DB

Open ML Notebook Sessions

Example ML Notebook

New ML Notebook

ML Analysis steps

Import OML python libraries

Create / join required data frames

Visualize input data

Split test / train data sets

Build Data science model

Run prediction & explain factors

Visualize charts

Store results

Deploy model

Data science PaaS Service (ADS)

Create Infrastructure

Create Data Science Project

Create Data Science Notebook

Open Jupyter Notebook

Example ML Notebook

New ML Notebook

ML Analysis steps

Import ADS / sklearn python libraries

Open dataset & Visualize in cell

Perform exploratory data analysis

Split test / train data sets

Evaluate models and explain

Visualize features & hyperparameters

Rank and visualize models

Score models (works with sklearn)

Save notebook

Deactivate Project

Algorithms – An overview

Classification / Regression & Ranking	Predicts target variable containing 2 (binary) or more (multi-class) category values	
	❖ *Decision Tree*	*- Generates human interpretable rules, primarily used for segmentation*
	❖ *Logistic regression / GLM*	*- Text semantic analysis and knowledge discovery*
	❖ *Naïve Bayes*	*- conditional probabilities assumes*
	❖ *Neural Network*	*- for most research deep learning applications for noisy/complex data with multiple hidden layers*
	❖ *Random Forest*	*- Extension of decision tree. More accurate prediction on larger data sets (i,e drug sensitivity)*
	❖ *Support vector machine*	*- linear/non-linear problems with multiple solvers & outlier detection. Popular in neuro-imaging area*
	❖ *Extreme gradient boosting*	*- Scalable implementation of XGBoost algorithm supporting tree / linear models*
	❖ *Stepwise regression*	*- best set of predictors (linear model) – forward/backward propagation and alternate direction*
Attribute Importance	Supervised and un-supervised ranking of variables to improve model quality	
	❖ *CUR Decomposition*	*- low rank SVD approach for ranking attribute importance as un-supervised method*
	❖ *Expectation maximization*	*- supports un-supervised variable ranking and pairwise dependency estimates*
	❖ *Minimum description length*	*- Most important variables for classification and regression*
Clustering	Group or segment cases into hierarchical clusters producing probabilities, rules, and statistics	
	❖ *K-means*	*- sparsity optimizations, outlier analysis, specified k clusters*
	❖ *Orthogonal partitioning*	*- discovers natural clusters up to max # specified and density-based*
	❖ *Expectation maximization*	*- Model search, overfitting protection and high quality probabilistic estimates*
Clustering	Derive new values where all Input variables considered to generate reduced set of variables	
	❖ *Explicit semantic analysis*	*- Text categorization & topic labels; semantic similarity estimates*
	❖ *Non-negative matrix factorization*	*- derives features based on non-negative linear combinations for feature interoperability*
	❖ *Principal component analysis (PCA)*	*- Works well on un-correlated variables with max variance*
	❖ *Singular value decomposition (SVD)*	*- Multiple solvers for narrow as well as wide data*
Anomaly detection Time series Association Rules & Row importance	❖ *One-class-SVM – (Anomaly detection)*	*- special solvers that do not use a target (linear/non-linear)*
	❖ *MSET-SPRT – (Anomaly detection)*	*- Process monitoring to detect linear/non-linear anomalies*
	❖ *Exponential smoothing – (Time series)*	*- Predict sequential numeric data using number or date/time columns*
	❖ *Apriori - (Association Rules)*	*- Market basic analysis using transactional or 2D representation of frequently occurring patterns*
	❖ *CUR Decomposition-(Row importance)*	*– Unsupervised ranking of rows – supports low-rank SVD based approach*

Guidance for Researchers

Autonomous & Oracle Machine Learning (OML)

Use

- When you need all your data inside database in a single place in cloud
- When you do not need to move large volumes of data & store in cloud
- When you need to import / export data from database within cloud fast
- When you prefer to work on both SQL and python interface on the notebooks (%SQL and %py)
- When you need to use Pre-built Data cleaning algorithms for data preparation
- When you need a rich support of popular data science algorithms with examples

Data Science Platform & Accelerated Data Science (ADS)

Use

- When data is distributed and are mostly in files
- When model evaluation and scoring is key
- When comparative model evaluation from sklearn / ADS providers are important
- When multiple researchers work on similar or same projects
- When you wish to leverage existing jupyter notebook and customize conda environments
- When you import and integrate existing shell / batch scripts for data preparation

All-in-one Data Science Image

Use

- When you are okay using latest python and CUDA & TensorFlow versions
- When you prefer to change GPU / CPU shapes often to test performance
- When you have custom ML/DL frameworks and libraries to load
- When cloud credits are at a premium and you need to terminate instances quickly
- When you need to import / export image between Oracle cloud / on-premisefrequently
- When you need to scale and test workloads on High end BM GPU shapes

Oracle for Research Images

Use

- When you prefer independence and start from a smaller conda sandbox environment
- When you are okay with sorting CUDA/Tensorflow version compatibilities
- When you wish separation of CPU / GPU Tensorflow versions or Ubuntu specific images

ORACLE
for Research

TECH TALK:

Data Science for Researchers

Questions, Answers & Discussion

ORACLE
for Research

TECH TALK:

Data Science for Researchers

Questions? Comments? Feedback?

Contact us!

Website: oracle.com/oracle-for-research/
Github: github.com/OracleforResearch
Twitter: @OracleResearch
Email: OracleForResearchTech_ww@oracle.com

Next Tech Talk: Mar 19th 2021

References

- Getting started : https://docs.oracle.com/en-us/iaas/data-science/data-science-tutorial/get-started.htm#create_notebook
- Conda Environments : https://docs.oracle.com/en-us/iaas/data-science/using/conda_understand_environments.htm
- Live Labs : https://apexapps.oracle.com/pls/apex/dbpm/r/livelabs/view-workshop?p180_id=673
- ADS : https://docs.oracle.com/en-us/iaas/tools/ads-sdk/latest/index.html#