MARKLOGIC WORLD 19

Accelerating Machine Learning with Smart Data Curation and GPUs

Anthony Roach
Senior Product Manager

Imran Chaudhri
Director, Industry Solutions

Biju George
Senior Solutions Engineer

Yangwei Liu
Software Engineer

“The sculpture is already complete within the marble block, before I start my work.

It is already there, I just have to chisel away the superfluous material.”

— MICHELANGELO

Session Agenda

- **Machine Learning Overview**
- **The 80/20 Rule and the Data Scientist**
- **MarkLogic Embedded Machine Learning**
- **Machine Learning in the MarkLogic and Data Hub**
- **Demo 1: The Fast, Secure Pipe**
- **Demo 2: In-Database Machine Learning**
- **Summary**
- **What's Next…**

Machine Learning

- Artificial Intelligence
  - Any technique which enables computers to mimic human behavior
- Machine Learning
  - Subset of AI techniques which use statistical methods to improve with experience
- Deep Learning
  - Subset of Machine Learning which makes the computation of neural networks feasible

Machine Learning is Pattern Recognition

Identifying non-local, non-linear relationships in a complex feature space

- A machine learning model is a representation of those relationships
- Models perform inference – prediction or classification
- Accuracy improvements
- Your business entities can be modelled with machine learning

What's making this possible now

Data

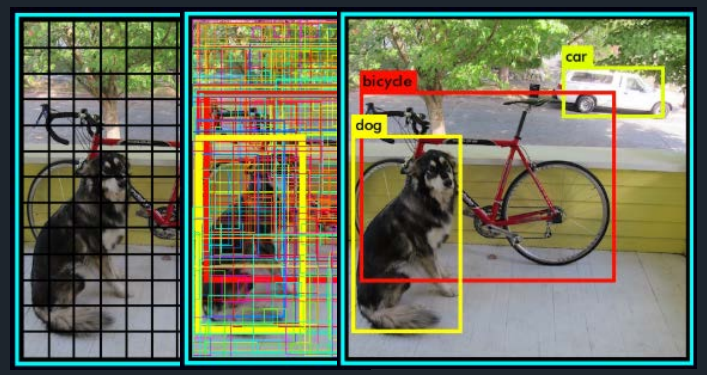
- Explosion of data
- Data Integration

"The biggest obstacle to using advanced data analysis isn't skill base or technology; it's plain old access to the data"

Ed Wilder-James
Harvard Business Review

Frameworks

- Frameworks are free
- Algorithms improving



Processing Power

- TFLOPS required for training
- GPUs bring power to the people

GPUs

DGX-1 with Tesla V100

8X GPU Server

CPU-only Server

Workload: ResNet50, 90 epochs

Amazon EC2 G2 Instance Family

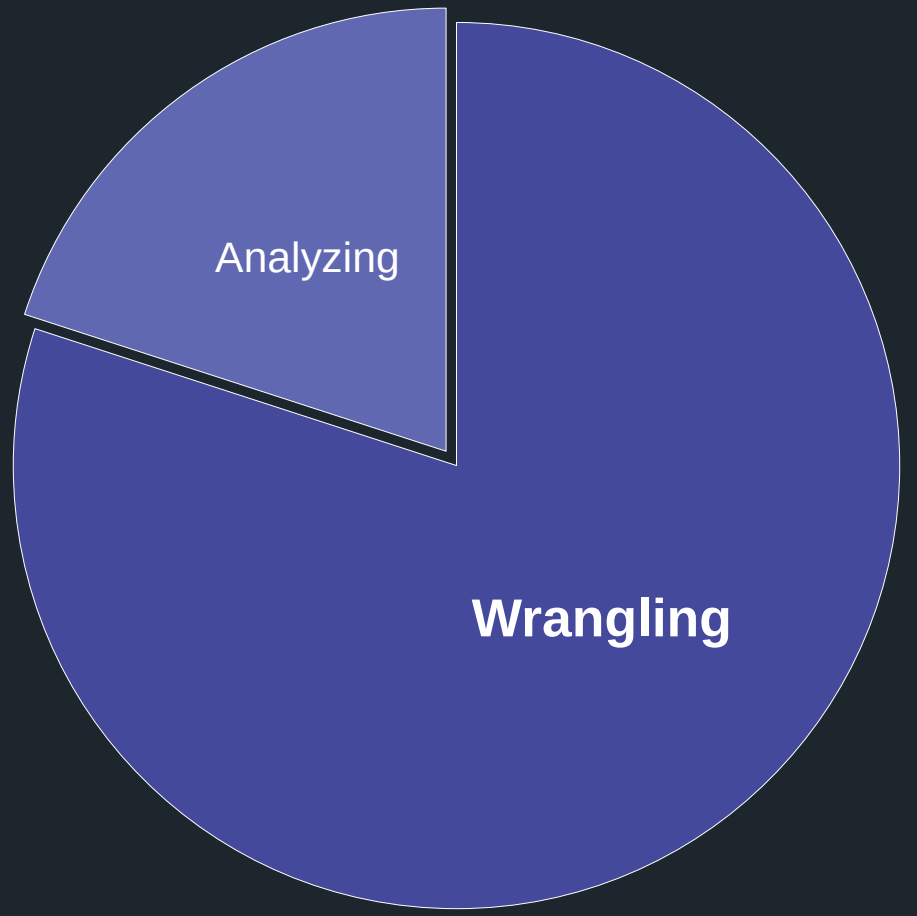
Type	NVIDIA GPU	vCPU	Mem (GiB)	SSD Storage (GB)
g2.2xlarge	1 x GK104	8	15	60
g2.4xlarge	2 x GK104	16	30	120
g2.8xlarge	4 x GK104	32	60	240

- Intel Xeon E5-2670 (Sandy Bridge)
- 1,536 CUDA cores & 4GB of video memory per GPU
- Hardware video encoder: 8x 720p or 4x 1080p @ 30fps
- Support for low-latency frame capture and encoding

amazon web services

The 80/20 rule

Most data scientists spend only 20 percent of their time on actual data analysis and 80 percent wrangling data.



Data Scientist

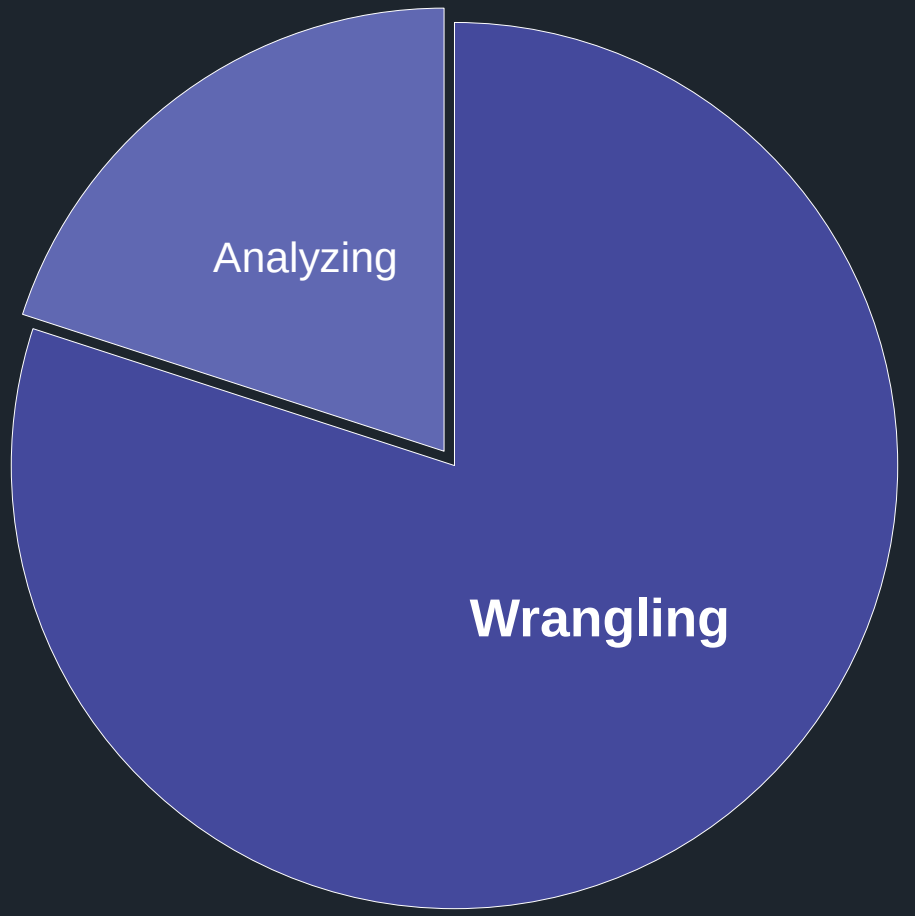
A statistician who lives in San Francisco

A device for turning coffee and data into better decisions

Someone who is better at statistics than any software engineer and better at software engineering than any statistician

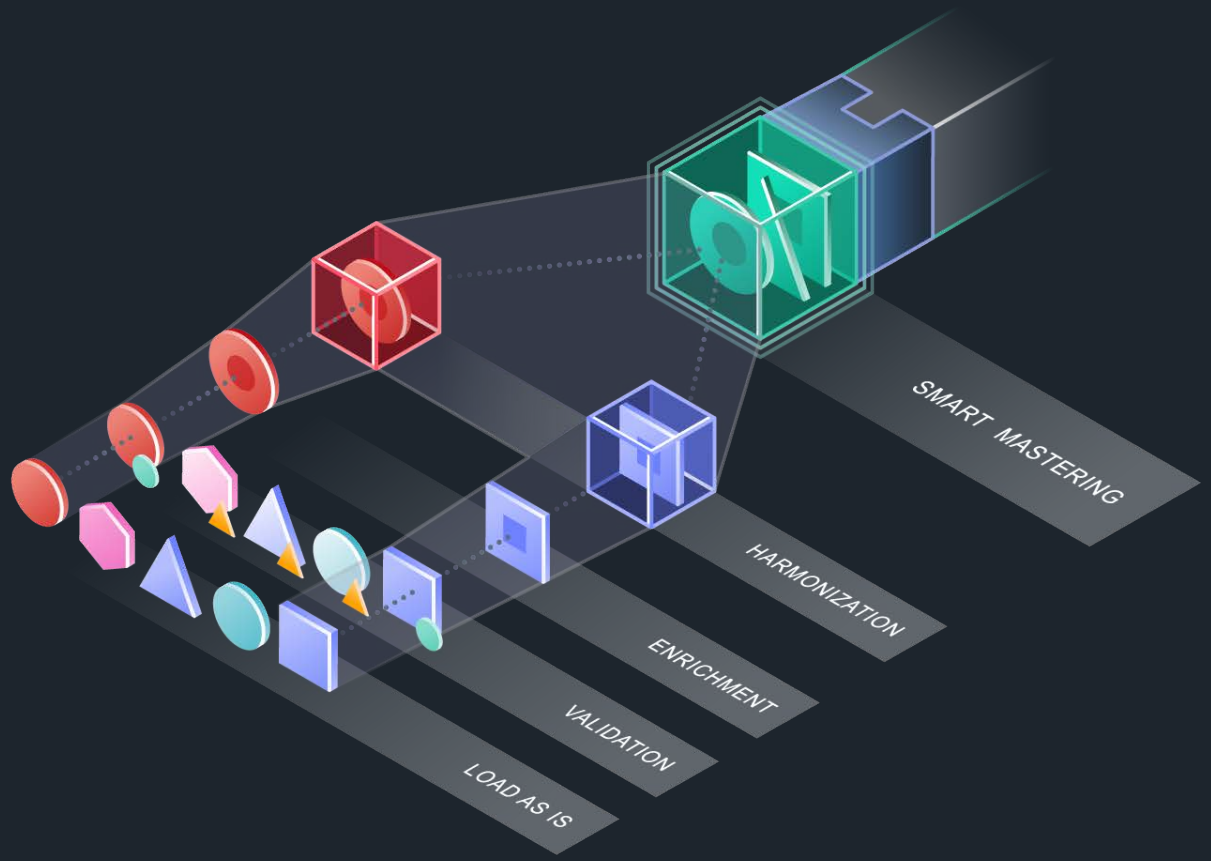
The 80/20 rule

Most data scientists spend only 20 percent of their time on actual data analysis and 80 percent wrangling data.



Data Wrangling

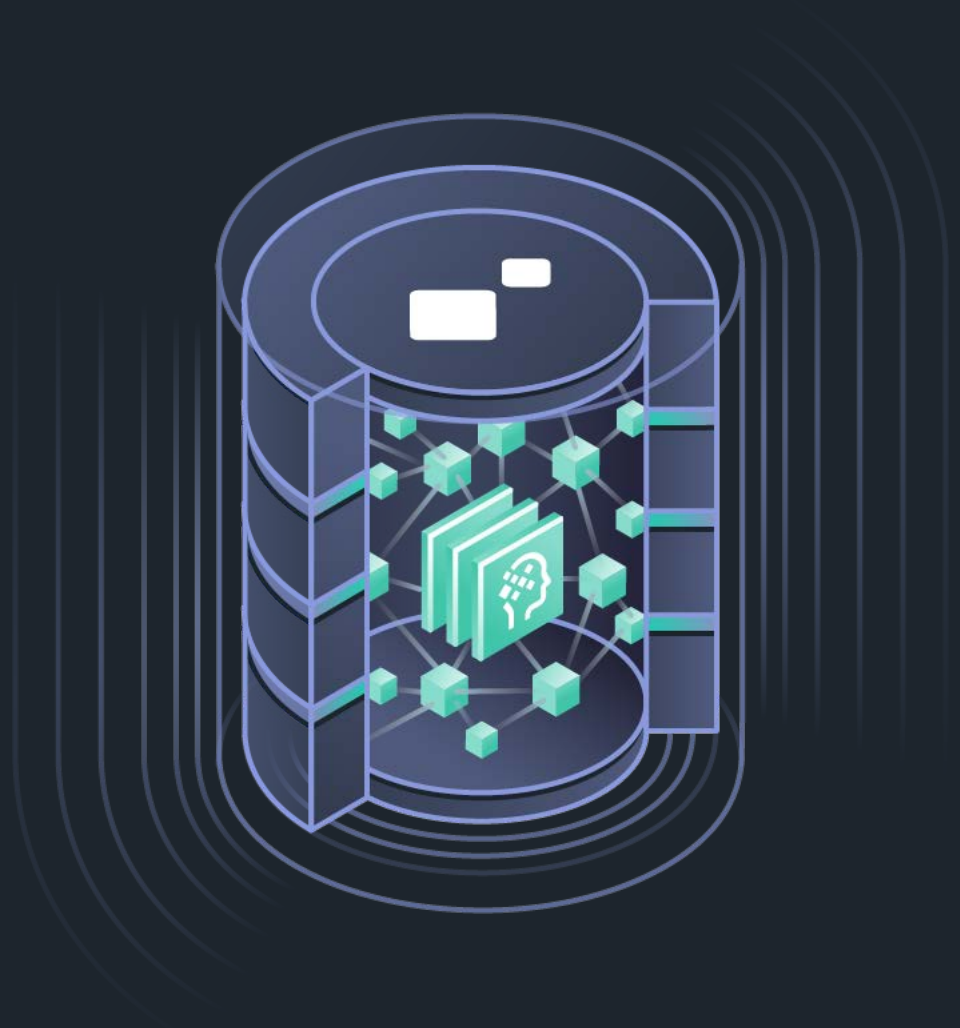
“the process of transforming and mapping data from one ‘raw’ data form into another ... with the intent of making it more appropriate and valuable for a variety of downstream purposes”



Curated Data

- Harmonized
- Mastered
- Reference data
- Enrichment
- Provenance and lineage

NEW



MarkLogic Embedded Machine Learning

Embedded Machine Learning

MarkLogic 10

- Microsoft CNTK libraries embedded in MarkLogic
  - Enterprise-grade, high-performance, fully-featured
  - Exposed through JavaScript and XQuery
- CUDA libraries embedded in MarkLogic
  - GPU acceleration (NVidia cards only)
  - Initially leveraged by CNTK

ONNX – Open Neural Network Exchange Format

- Embedded with CNTK
- Developed by Microsoft, Facebook, AWS
- Move models between state-of-the-art stacks
- High performance, hardware optimized
- Remove ecosystem lock-in

Machine Learning in MarkLogic

- Smarter Mastering
- Autonomous elasticity
- Predictive query optimization
- Proactive curation and governance
- Roll your own

MarkLogic

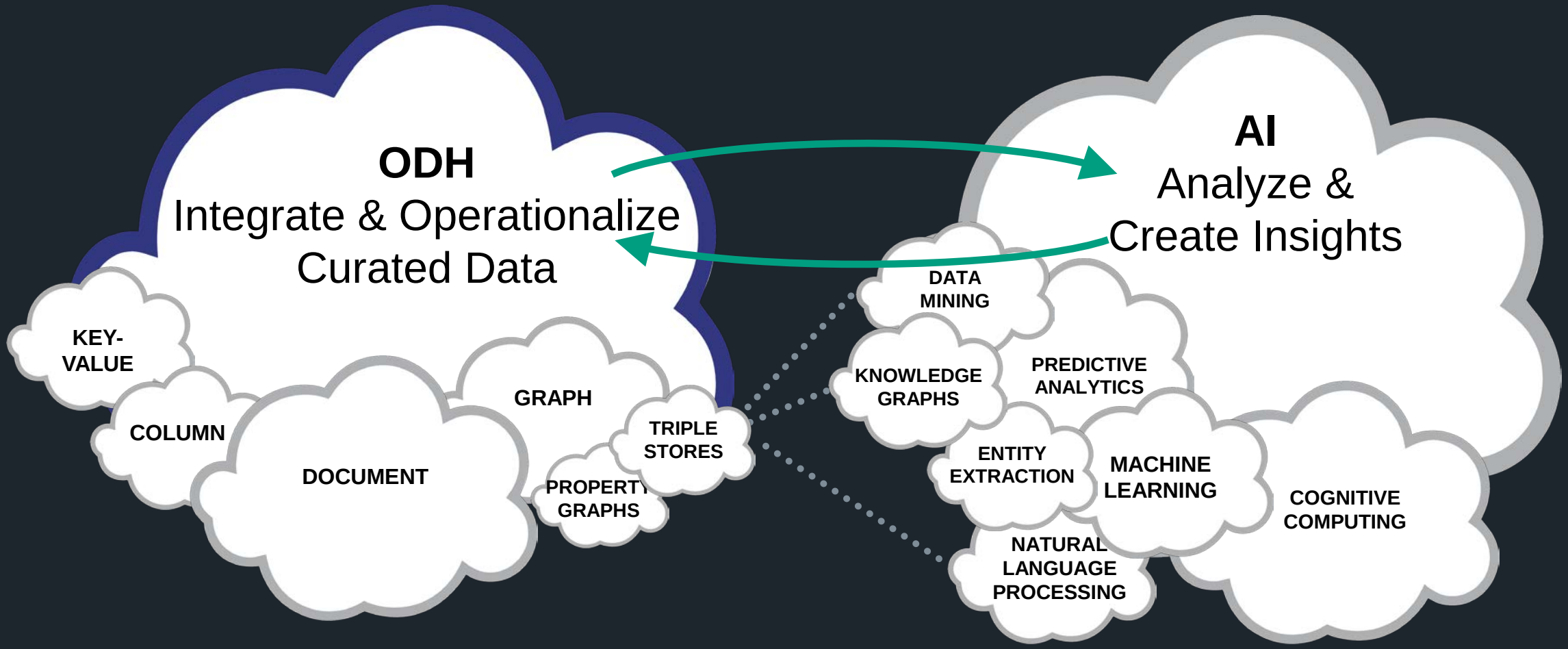
Use Case 1

High-Performance Connector

Imran Chaudhri
Director, Industry Solutions

Biju George
Senior Solutions Engineer

ODH and Data Data Analytics Roles



Data Science Learning Objective

- ***Question***

  ‘What is the projected order quantity for each product for next three years?’

- ***Algorithm***

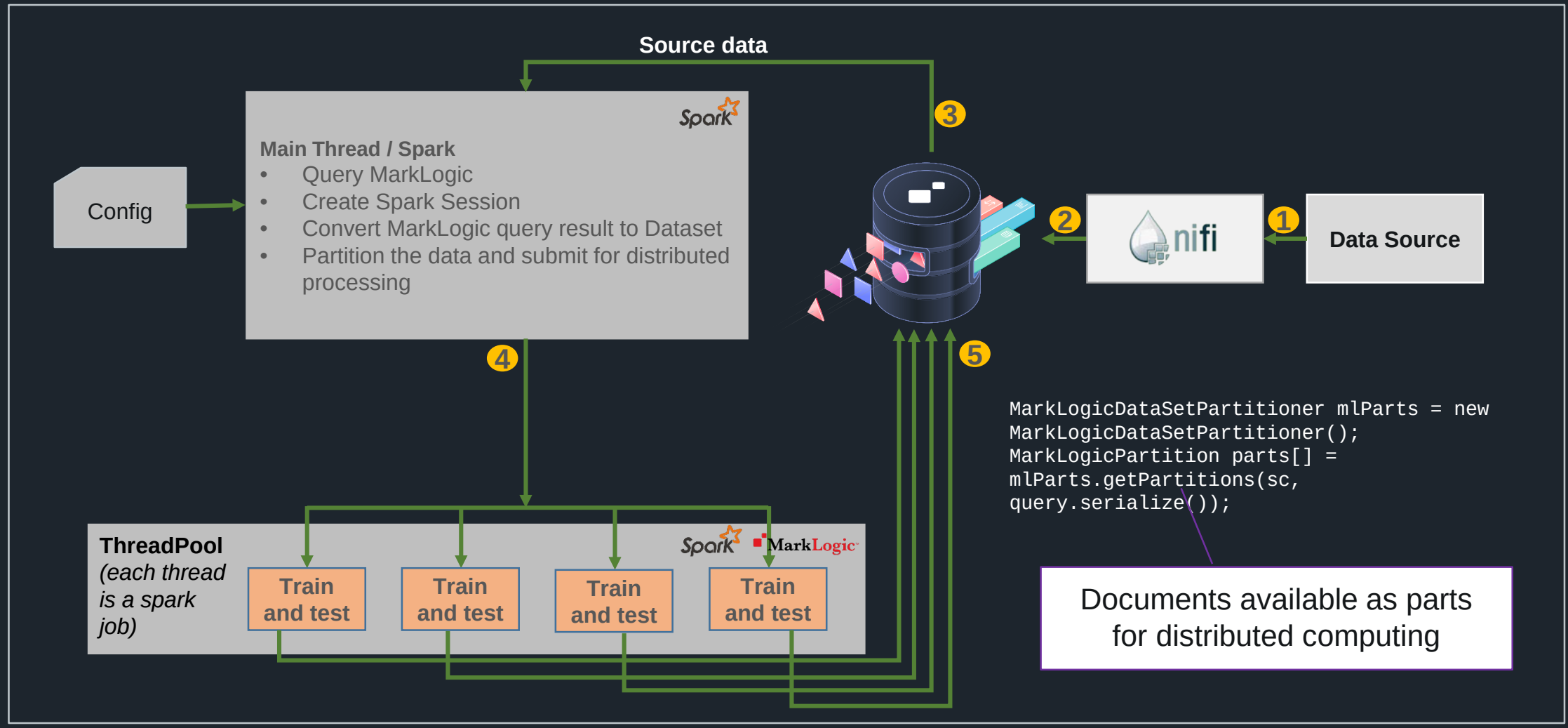
  ‘Linear Regression’

- ***Technology***

  - *MarkLogic Server Essential Enterprise 9.0-9+*
  - *Spark 2.3.2*
  - *Java 1.8*
  - *Java dependencies (Refer pom.xml)*

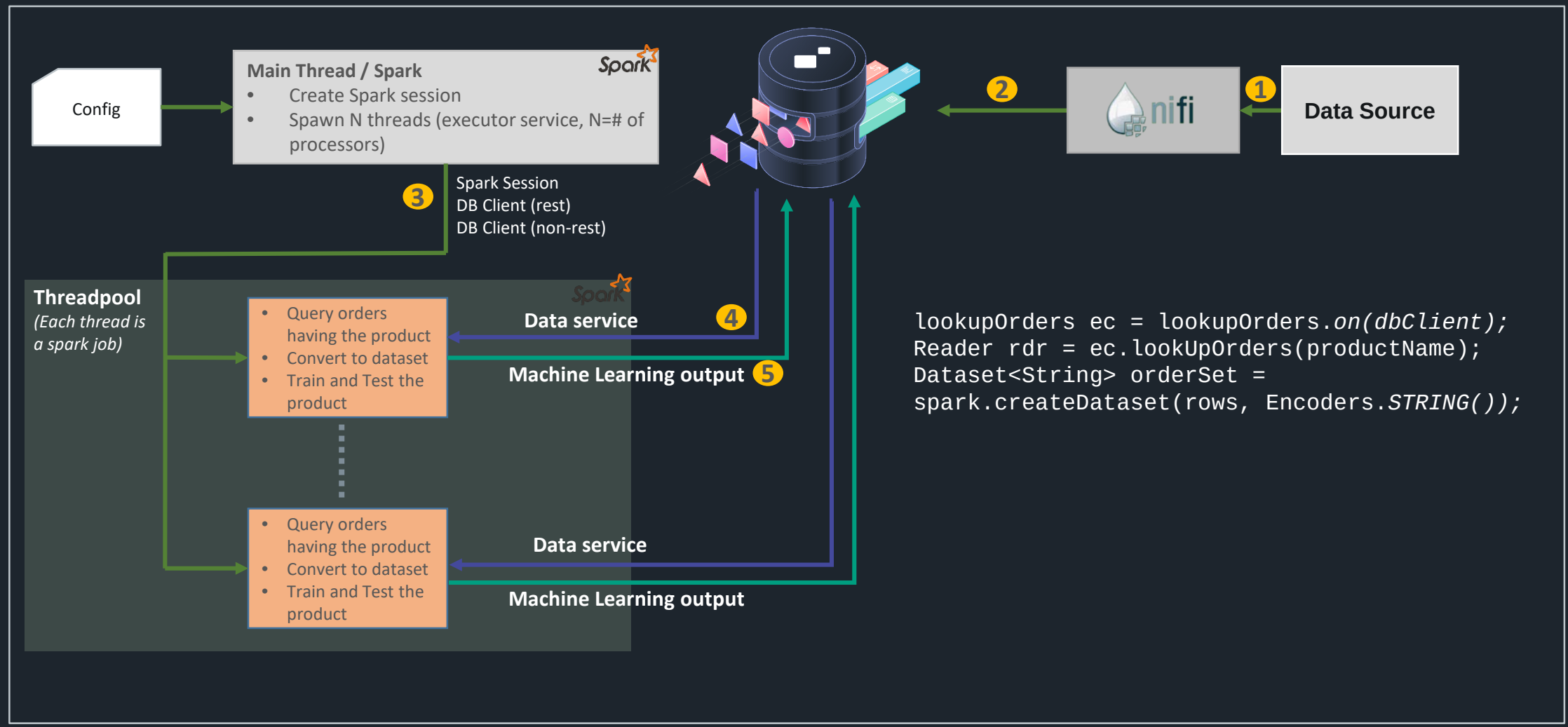
1 – Partition MarkLogic Data

MarkLogic data partitioned for Spark by central orchestrator using DMSDK.



2 - Data Services Approach

Each Spark worker can directly query MarkLogic for its own data using Data Services.

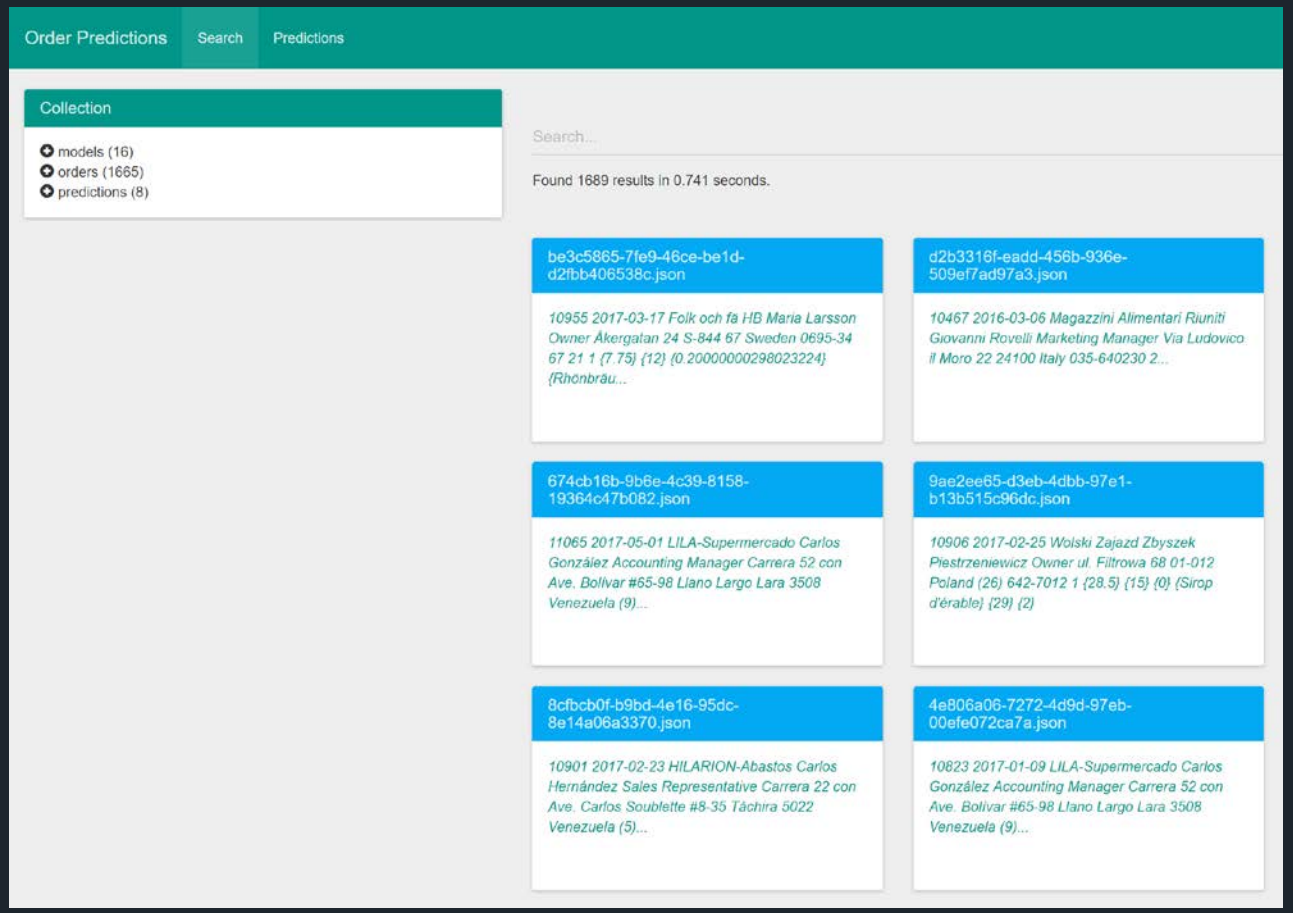



```
lookupOrders ec = lookupOrders.on(dbClient);
Reader rdr = ec.lookUpOrders(productName);
Dataset<String> orderSet =
spark.createDataset(rows, Encoders.STRING());
```

One place for all data for data scientists

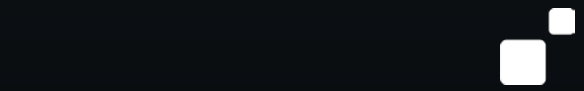
Training Data, Training
Models and training results

- All available real-time in the same database for the data scientists and business users
- 360º of the all the data



Demo

High-Performance Connector



MARKLOGIC WORLD 19

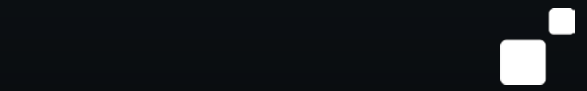
MarkLogic 10 Demo

Embedded Machine Learning

Yangwei Liu
Software Engineer

Demo

Embedded Machine Learning



Summary

- It’s all about the data
- Uncover hidden insights
- Supercharge your data scientists
- Build a smarter Data Hub

What's next

- All MarkLogic products will be augmented by machine learning
  - Data Hub capabilities
  - Database performance
  - Service manageability
- MarkLogic as an integral part of a data science toolchain
  - Best source for machine learning data
  - Best platform for integrating machine learning into your enterprise