

Machine Learning for Data Management Systems

Introduction

Amol Deshpande
Jan 26, 2023

Outline

- Motivation
- Course Goals and Focus Areas
- Intros
- Overview of Data Management Systems
- History of Database Automation

Promise of Big Data



- Explosion of data, in pretty much every domain
 - Sensing devices and sensor networks that can monitor everything 24/7 from temperature to pollution to vital signs
 - Increasingly sophisticated smart phones
 - Internet, social networks make it easy to publish data
 - Scientific experiments and simulations → astronomical data volumes
 - Genome/health data
 - Internet of Things, Smart wearables
 - ...

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021

INFRASTRUCTURE

STORAGE
Amazon S3, Google Cloud Storage, IBM Cloud Storage, Pure Storage, Veeva, NetScout Systems, Cohesity, Vast, DataEye, Qumulo

HADOOP
Cloudera, Amazon EMR, Databricks, AWS EMR, Hadoop Distribution, Pivotal, Jethro, Cloud Platform, OGM, IBM InfoSphere

DATA LAKES
Databricks, Amazon Data Lake Storage, Dremio, AWS Lake Formation, Firebolt, Hazelcast, Ix, Materialize

DATA WAREHOUSES
Amazon Redshift, Snowflake, Google BigQuery, Microsoft Azure Synapse Analytics, Oracle Exadata, IBM Db2 Warehouse, Amazon Redshift, Snowflake, Google BigQuery, Microsoft Azure Synapse Analytics, Oracle Exadata, IBM Db2 Warehouse

STREAMING / IN-MEMORY
Amazon Kinesis, Databricks, Apache Flink, Amazon Kinesis, Databricks, Apache Flink, Amazon Kinesis, Databricks, Apache Flink

DBMS
Oracle, Microsoft SQL Server, IBM Db2, SAP HANA, Amazon RDS, Amazon Aurora, Oracle, Microsoft SQL Server, IBM Db2, SAP HANA, Amazon RDS, Amazon Aurora

NOSQL DATABASES
Amazon DynamoDB, Amazon ElastiCache, Amazon DocumentDB, SAP HANA, Amazon DynamoDB, Amazon ElastiCache, Amazon DocumentDB

NEWSQL DATABASES
CockroachDB, YugabyteDB, Amazon Aurora, SAP HANA, CockroachDB, YugabyteDB, Amazon Aurora, SAP HANA

REAL TIME DATABASES
Amazon Kinesis, Amazon Redshift, Amazon ElastiCache, Amazon Kinesis, Amazon Redshift, Amazon ElastiCache

GRAPH DBS
Amazon Neptune, Oracle, SAP HANA, Amazon Neptune, Oracle, SAP HANA

MPP DBS
Teradata, Vertica, Amazon Redshift, Amazon ElastiCache, Amazon DocumentDB, Amazon Redshift, Amazon ElastiCache, Amazon DocumentDB

ETL / ELT / DATA TRANSFORMATION
Talend, Alteryx, Informatica, Amazon Redshift, Amazon ElastiCache, Amazon DocumentDB, Talend, Alteryx, Informatica, Amazon Redshift, Amazon ElastiCache, Amazon DocumentDB

REVERSE ETL
Census, Grouparoo, PubliStack, Pipedrive, Pipedrive, Pipedrive

DATA INTEGRATION
MuleSoft, Tealium, Inoplog, MuleSoft, Tealium, Inoplog

DATA CATALOG AND DISCOVERY
Metaphor, Atlan, Data.world, Supergrain, Metaphor, Atlan, Data.world, Supergrain

METRICS STORE
GoodData, Splunk, GoodData, Splunk

LOG ANALYTICS
Splunk, Sumologic, Splunk, Sumologic

COMPUTER VISION
Microsoft Azure, Cloud Vision API, Amazon Rekognition, Microsoft Azure, Cloud Vision API, Amazon Rekognition

SPEECH
Siri, Amazon Alexa, Microsoft Azure, Cloud Speech, Amazon Rekognition, Siri, Amazon Alexa, Microsoft Azure, Cloud Speech, Amazon Rekognition

NLP
Google Natural Language API, Amazon Comprehend, Microsoft Azure, Text Analytics, Google Natural Language API, Amazon Comprehend, Microsoft Azure, Text Analytics

SYNTHETIC MEDIA
DeepBrain AI, D-ID, Synthesia, D-ID, Synthesia

ADVERTISING
Xandr MediaMath, Criteo, IAS, Albert, Gungum, Xandr MediaMath, Criteo, IAS, Albert, Gungum

EDUCATION
Knewton, K12, Knewton, K12

REAL ESTATE
Redfin, VTS, Opener, Orchard, Redfin, VTS, Opener, Orchard

GOVT / INTELLIGENCE
Palantir, OpenView, Palantir, OpenView

COMMERCE
Stitch Fix, Shein, Faire, Affirm, Mercado, Root, Automia, Stitch Fix, Shein, Faire, Affirm, Mercado, Root, Automia

FINANCE - LENDING
Affirm, Mercado, Root, Automia, Affirm, Mercado, Root, Automia

INSURANCE
Root, Automia, Root, Automia

HEALTHCARE
Flatiron, Xyris, Metabion, Flatiron, Xyris, Metabion

LIFE SCIENCES
Flatiron, Xyris, Metabion, Flatiron, Xyris, Metabion

TRANSPORTATION
Uber, Tesla, Cruise, Aptoiv, Uber, Tesla, Cruise, Aptoiv

AGRICULTURE
Granular, Granular

INDUSTRIAL
AVEVA, Siemens, AVEVA, Siemens

OTHER
ByteDance, Stem, ByteDance, Stem

PRIVACY & SECURITY
Very Good Security, Privabera, Crystal, Casp Privacy, Privacy Dynamics, Very Good Security, Privabera, Crystal, Casp Privacy, Privacy Dynamics

DATA OBSERVABILITY
Datastack, Monte Carlo, Datastack, Monte Carlo

MGMT / MONITORING
Cribl, Moogsoft, Chronosphere, Cribl, Moogsoft, Chronosphere

SERVERLESS
Amazon Lambda, AWS Fargate, Amazon Lambda, AWS Fargate

CLUSTER SVCS
Amazon EMR, Amazon ElastiCache, Amazon DocumentDB, Amazon EMR, Amazon ElastiCache, Amazon DocumentDB

ANALYTICS

BI PLATFORMS
Looker, Tableau, Power BI, Microsoft Power BI, Tableau, Power BI

VISUALIZATION
Tableau, Power BI, Microsoft Power BI, Tableau, Power BI

DATA ANALYST PLATFORMS
Microsoft, Pentaho, Alteryx, Microsoft, Pentaho, Alteryx

AUGMENTED ANALYTICS
ThoughtSpot, Anodot, Outlier, ThoughtSpot, Anodot, Outlier

DATA CATALOG AND DISCOVERY
Metaphor, Atlan, Data.world, Supergrain, Metaphor, Atlan, Data.world, Supergrain

METRICS STORE
GoodData, Splunk, GoodData, Splunk

LOG ANALYTICS
Splunk, Sumologic, Splunk, Sumologic

QUERY ENGINE
Dremio, Starburst, Ahana, Dremio, Starburst, Ahana

SEARCH
Elasticsearch, Amazon Kendra, Oracle Endeca, Amazon CloudSearch, Elasticsearch, Amazon Kendra, Oracle Endeca, Amazon CloudSearch

MACHINE LEARNING & ARTIFICIAL INTELLIGENCE

DATA SCIENCE NOTEBOOKS
JupyterLab, Binder, Colab, JupyterLab, Binder, Colab

DATA SCIENCE PLATFORMS
DataRobot, Dataiku, DataRobot, Dataiku

ML PLATFORMS
DataRobot, Dataiku, DataRobot, Dataiku

DATA GENERATION & LABELING
Scale AI, Upwork, Appen, Scale AI, Upwork, Appen

MODEL BUILDING
Weights & Biases, Weights & Biases

FEATURE STORE
Feast, Featurehub, Feast, Featurehub

DEPLOYMENT & PRODUCTION
DataRobot, Dataiku, DataRobot, Dataiku

MODEL MONITORING & OBSERVABILITY
Arthor, WhyLabs, Arthor, WhyLabs

COMPUTER VISION
Microsoft Azure, Cloud Vision API, Amazon Rekognition, Microsoft Azure, Cloud Vision API, Amazon Rekognition

SPEECH
Siri, Amazon Alexa, Microsoft Azure, Cloud Speech, Amazon Rekognition, Siri, Amazon Alexa, Microsoft Azure, Cloud Speech, Amazon Rekognition

NLP
Google Natural Language API, Amazon Comprehend, Microsoft Azure, Text Analytics, Google Natural Language API, Amazon Comprehend, Microsoft Azure, Text Analytics

SYNTHETIC MEDIA
DeepBrain AI, D-ID, Synthesia, D-ID, Synthesia

HORIZONTAL AI
IBM Watson, OpenAI, DeepMind, IBM Watson, OpenAI, DeepMind

GPU DBS & CLOUD
Kinetic, Paperspace, Kinetic, Paperspace

AI HARDWARE
Google TPU, ARM, Intel AI, Google TPU, ARM, Intel AI

APPLICATIONS - ENTERPRISE

SALES
Salesforce, Salesforce, Salesforce

MARKETING B2B
App Annie, Sense, App Annie, Sense

MARKETING - B2C
Google Analytics, Tealium, ActionIQ, Google Analytics, Tealium, ActionIQ

CUSTOMER EXPERIENCE / SERVICE
Qualtrics, Zendesk, SurveyMonkey, Qualtrics, Zendesk, SurveyMonkey

HUMAN CAPITAL
Workday, Workday, Workday

LEGAL
Ravel, Disco, Ravel, Disco

REGTECH & COMPLIANCE
Bigo, Regtech, Bigo, Regtech

FINANCE
Anaplan, Zelus, Finance, Anaplan, Zelus, Finance

AUTOMATION & RPA
UiPath, Celonis, Zinnov, UiPath, Celonis, Zinnov

SECURITY
Tanium, Palo Alto Networks, Palo Alto Networks, Palo Alto Networks

ADVERTISING
Xandr MediaMath, Criteo, IAS, Albert, Gungum, Xandr MediaMath, Criteo, IAS, Albert, Gungum

EDUCATION
Knewton, K12, Knewton, K12

REAL ESTATE
Redfin, VTS, Opener, Orchard, Redfin, VTS, Opener, Orchard

GOVT / INTELLIGENCE
Palantir, OpenView, Palantir, OpenView

COMMERCE
Stitch Fix, Shein, Faire, Affirm, Mercado, Root, Automia, Stitch Fix, Shein, Faire, Affirm, Mercado, Root, Automia

FINANCE - LENDING
Affirm, Mercado, Root, Automia, Affirm, Mercado, Root, Automia

INSURANCE
Root, Automia, Root, Automia

HEALTHCARE
Flatiron, Xyris, Metabion, Flatiron, Xyris, Metabion

LIFE SCIENCES
Flatiron, Xyris, Metabion, Flatiron, Xyris, Metabion

TRANSPORTATION
Uber, Tesla, Cruise, Aptoiv, Uber, Tesla, Cruise, Aptoiv

AGRICULTURE
Granular, Granular

INDUSTRIAL
AVEVA, Siemens, AVEVA, Siemens

OTHER
ByteDance, Stem, ByteDance, Stem

OPEN SOURCE

FRAMEWORKS
TensorFlow, PyTorch, TensorFlow, PyTorch

FORMAT
JSON, XML, CSV, JSON, XML, CSV

QUERY / DATA FLOW
Apache Airflow, Apache Airflow

DATA ACCESS
Uber Databook, Amazon Athena, Uber Databook, Amazon Athena

DATABASES
MySQL, PostgreSQL, MySQL, PostgreSQL

ORCHESTRATION
Apache Airflow, Apache Airflow

INFRA-STRUCTURE
Kubernetes, Docker, Kubernetes, Docker

DATA OPS
Marquez, Marquez

STREAMING & MESSAGING
Apache Kafka, Apache Kafka

STAT TOOLS & LANGUAGES
R, Python, R, Python

ML OPS & INFRA
MLflow, MLflow

AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow, PyTorch, TensorFlow, PyTorch

SEARCH
Elasticsearch, Elasticsearch

LOGGING & MONITORING
Datadog, Datadog

VISUALIZATION
Tableau, Power BI, Tableau, Power BI

COLLABORATION
Slack, Slack

SECURITY
OpenStack, OpenStack

DATA SOURCES & APIS

DATA MARKETPLACES & DISCOVERY
Bloomberg, Thomson Reuters, Dow Jones, Quandl, Bloomberg, Thomson Reuters, Dow Jones, Quandl

FINANCIAL & ECONOMIC DATA
Bloomberg, Thomson Reuters, Dow Jones, Quandl, Bloomberg, Thomson Reuters, Dow Jones, Quandl

AIR / SPACE / SEA
Airbus, Boeing, Airbus, Boeing

PEOPLE / ENTITIES
ZoomInfo, Axciom, People, ZoomInfo, Axciom, People

LOCATION INTELLIGENCE
FourSquare, Mapbox, Esri, FourSquare, Mapbox, Esri

OTHER
Data.gov, Data.gov

DATA RESOURCES

DATA SERVICES
Kaggle, Kaggle

INCUBATORS & SCHOOLS
Pluralsight, General Assembly, DataCamp, Pluralsight, General Assembly, DataCamp

RESEARCH
OpenAI, Google Research, Facebook Research, OpenAI, Google Research, Facebook Research

Backbone = Data Management Systems

- Need better and bigger data management systems to support these use cases
 - To handle the much much larger datasets (Volume)
 - To respond quickly to new data (Velocity)
 - To manage and query a wide variety of complex data types (Variety)
 - To properly reason about robustness and other issues (Veracity)

40 ZETTABYTES

(40 TRILLION GIGABYTES)
of data will be created by 2020, an increase of 300 times from 2005



It's estimated that **2.5 QUINTILLION BYTES**

(2.5 TRILLION GIGABYTES)
of data are created each day



Volume
SCALE OF DATA

6 BILLION PEOPLE
have cell phones



WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least **100 TERABYTES**
(100,000 GIGABYTES)
of data stored



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity.

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015, **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States.



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
(150 BILLION GIGABYTES)



By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

Variety
DIFFERENT FORMS OF DATA

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month



30 BILLION PIECES OF CONTENT are shared on Facebook every month



400 MILLION TWEETS are sent per day by about 200 million monthly active users



The New York Stock Exchange captures

1 TB OF TRADE INFORMATION during each trading session



Velocity
ANALYSIS OF STREAMING DATA



Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

— almost 2.5 connections per person on earth



1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions



Poor data quality costs the US economy around

\$3.1 TRILLION A YEAR



27% OF RESPONDENTS

Veracity
UNCERTAINTY OF DATA

in one survey were unsure of how much of their data was inaccurate

Modern Data Management Systems

- Much more complex than in the past
 - Deployed over 100's or 1000's of servers (or more)
 - With rapidly changing configurations
 - Often deployed through virtualization on clouds
 - Many combinations of hardware technologies
 - CPU/GPUS, Complex Cache Hierarchies, Direct Remote Memory Access, Fast interconnects, ...
 - Likely a mix of hardware with different characteristics in a single setup
 - Frequent changes, upgrades, etc.

Modern Data Management Systems

- Much more complex than in the past
 - Support many different types of data
 - JSON, Video, Timeseries, Audio, Text, Geospatial, ...
 - More complex query languages with many features
 - More operations, user-defined functions...
 - New query languages (e.g., Apache Spark, MongoDB QL)
 - Recently built-in support for ML training and inference
 - Corresponding increase in the software complexity
 - New types of join operators and indexes

Modern Data Management Systems

- Many more DMS today than in the past
 - Relational (SQL-based) database systems
 - Stream processing systems (focusing on streaming data)
 - Special-purpose data warehousing systems (most start from some RDBMS)
 - Batch analysis frameworks (like Hadoop, Pregel, Spark, ...)
 - Typically, data stored in distributed file systems
 - Key-value stores (like HBase, Cassandra, Redis, ...)
 - Basically, persistent distributed hash tables
 - Semi-structured/Document data stores (for XML/JSON query processing)
 - Graph databases
 - Data lakes (e.g., scientific data, machine learning data)

1. Tuning Data Management Systems

- Data management systems have many “knobs” (tuning params)
 - max #connections, shared memory, cache size, when to garbage collect, how often to run statistics, how to allocate memory across components, commit parameters, ...
 - PostgreSQL has about 170 knobs -- a small fraction with significant impact
 - Which materialized views to maintain, what indexes to use, what compression schemes (in data warehouses), window sizes (in streaming systems), what keys to use for partitioning data, how to partition, how many machines to use for query processing, ...
 - Most have significant impact on performance

1. Tuning Data Management Systems

- In the past, most decisions made by “DBAs”
- Much harder to “tune” or “configure” modern systems
 - Too many variables and too many combinations → Hard for humans to reason about
 - Lot of trial and error required, not feasible at the data volumes
 - Too many different systems → Hard to build up the experience
 - Environment variables changing too rapidly
- **Motivation 1: Build autonomous data management systems using ML**

2. “Learned” Components

- Many complex trade-offs when making design decisions in a database system
 - Different indexes better for different environments/different workloads
 - Same for storage layouts and other design decisions
- Motivation 2: Could we use modern ML techniques to design new self-adapting components, that can learn from the data/workload and automatically do the right thing for the given data/workload?

3. Workload Forecasting

- Better understanding of the future workload can help with planning through...
 - Allocating additional resources proactively rather than reactively
 - Exploiting different tradeoffs (e.g., using less memory per task if many tasks expected)
- Caveat: There must be patterns to be learned from
- Motivation 3: Incorporate forecasting algorithms to improve overall performance

4. Intra-query Adaptivity

- For complex queries/analysis tasks, things can change significantly during execution of a single query/task
 - Data characteristics may be very different than expected
 - Resources may fluctuate significantly during execution
- Motivation 4: Use ML techniques to adapt during the execution of a single query/task

5. Hard Planning/Optimization Problems

- Quite a few NP-Hard planning/optimization problems being solved in systems
 - (Query optimization) choosing a “query plan” given a complex query
 - Partitioning strategies in distributed systems, etc.
- Often need to be solved in presence of significant “uncertainties”
- Motivation 5: Could use of ML techniques provide different solutions to such problems? If yes, why?
 - Recent work on how deep learning could be used to “partially” solve hard combinatorial problems

6. User Interfaces/Interactions

- Natural language interfaces to querying (e.g., through conversion to SQL)
- Inferring user intent and responding accordingly with the right data/graphs
- Reducing the time to design schemas and build end-to-end applications
- Motivation 6: Using LLMs (large language models) and other such technologies to improve these facets

7. Miscellaneous

- Capturing correlations in the data for better estimation of query sizes (for optimization or approximate query processing)
- Synthetic data generation (e.g., to preserve privacy) through use of generative models
- Better dataset discovery and correlation in data lakes
- ...

Summary

- Modern data management systems are too complex to manage
- Many ways to incorporate ML techniques
 - to improve performance through forecasting and adapting
 - to reduce friction in user interactions
 - to obtain better optimization algorithms in face of uncertainty
 - ...
- Many other places where ML comes up in data management
- Also, much work on using database techniques to improve ML

Outline

- Motivation
- **Course Goals and Focus Areas**
- Intros
- Overview of Data Management Systems
- History of Database Automation

Course Goals

- Overarching goal: How to rearchitect modern data management systems to utilize advances in ML especially deep learning
 - Evolutionary (e.g., better forecasting), or
 - Revolutionary (entirely change how indexing or QO is done)
- More specifically:
 - Study the recent work on applying ML to data management systems
 - Reason about whether the use of ML is appropriate and why prior techniques can't be adapted
 - Think through the failure scenarios
 - Understand fundamental reasons (if any) why ML-based approach is superior
 - Explore other places where ML could help (especially LLMs)
 - Simplify data management systems through use of ML

Topics

- Learned Indexes and Storage Layouts
 - Improve performance of search and storage organization through learning
- Query Processing
 - Adaptive query operators, as well as adaptive query processing
- Query Optimization
 - Better estimations through capturing correlations, better search algorithms
- Natural Language to SQL
- Workload forecasting and resource management
- *May adjust as the semester goes on*

Approach

- Read 1-2 papers per class, mostly from database/systems conferences
 - I will try to provide the relevant background on the DB side
 - Coverage of ML techniques as required for the papers
 - May take breaks in between to cover some of the ML background in more depth
- Discuss each paper in the class
 - with approx. 45-minute presentation by one of you (will circulate sign up sheet)
 - Primary aim to discuss the papers deeply
 - Secondary goal to cover the broader topic of the paper but hard to do given how new the work is
- A few classes dedicated to broader discussions

Grading

- Paper Readings + class participation, etc. (20%)
 - Submissions through Gradescope (not Slack)
 - Due 11am of the day of the class
- Written Assignments/Final -- all individual (50%)
 - Spread throughout the semester -- will cover additional papers
 - Survey assignment: One written assignment will be doing a literature survey on one of the relevant topics and summarizing the recent work in that topic
 - One assignment on proposing a new idea in this space
- Research Project (30%)
 - Group research project

Other Logistics

- All submissions (paper critiques, assignments, project deliverables) through Gradescope
 - Will share all submitted critiques after the deadline
- Slack to be used for announcements/discussions
- Will try to move (at least some of) the classes to Iribe

Outline

- Motivation
- Course Goals and Focus Areas
- **Intros**
- Overview of Data Management Systems
- History of Database Automation

Next Steps...

- Sign up for Slack and Gradescope (not set up yet)
- Look out for the sign-up sheet for class presentations (starting the week after next)
- Readings for next week (more background)

- We will start with “Architecture of a Traditional Database System” in the next class