slides credit: Matthias Boehm

SCIENCE
PASSION
TECHNOLOGY

Data Integration and Large Scale Analysis
12 Distributed ML Systems

Shafaq Siddiqi

Graz University of Technology, Austria

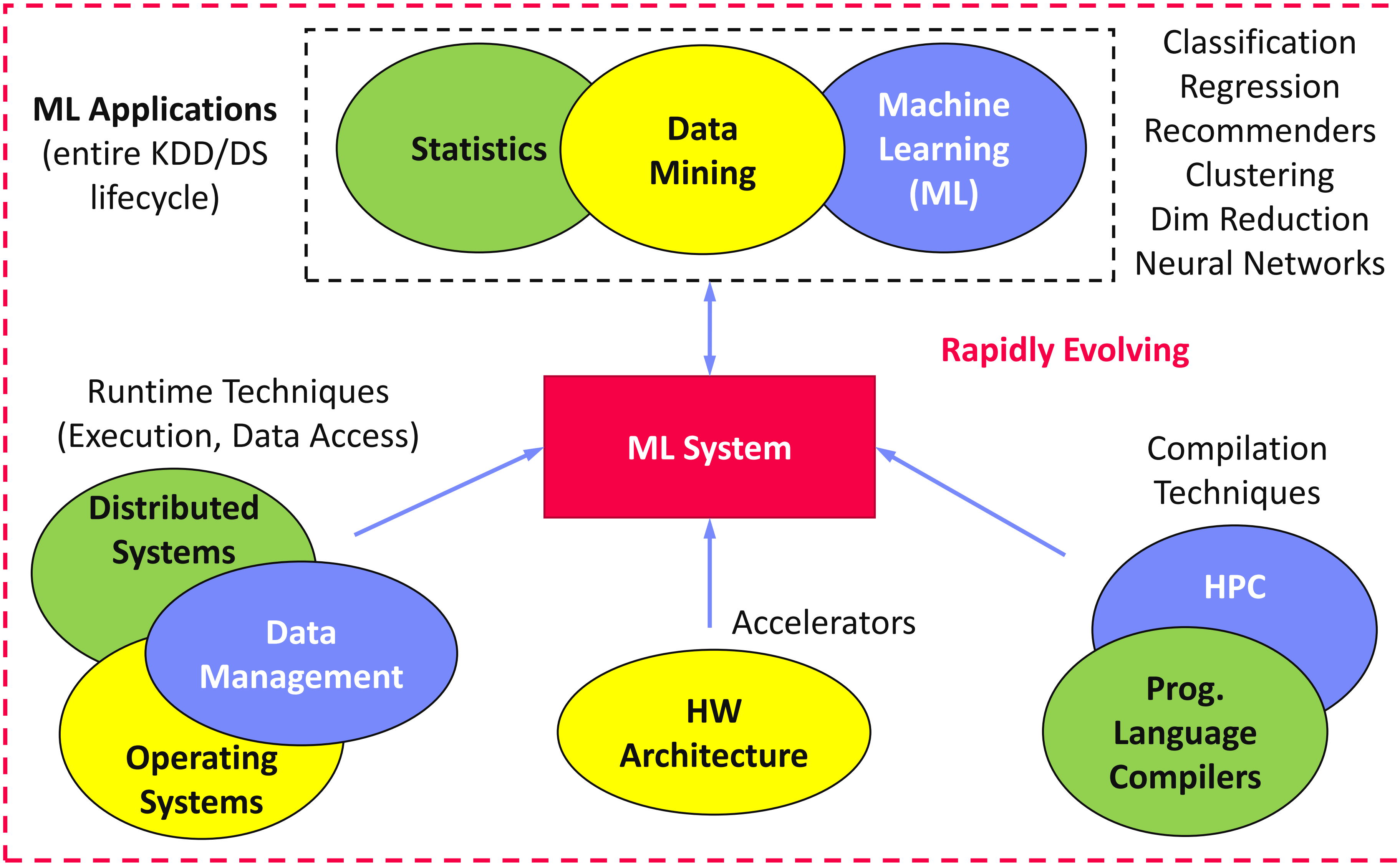
Last update: Jan 19, 2024

Agenda

- **Landscape of ML Systems**
- **Distributed Parameter Servers**
- **Large Language Models**
- **Q&A and Exam Preparation**

Landscape of ML Systems

What is an ML System?

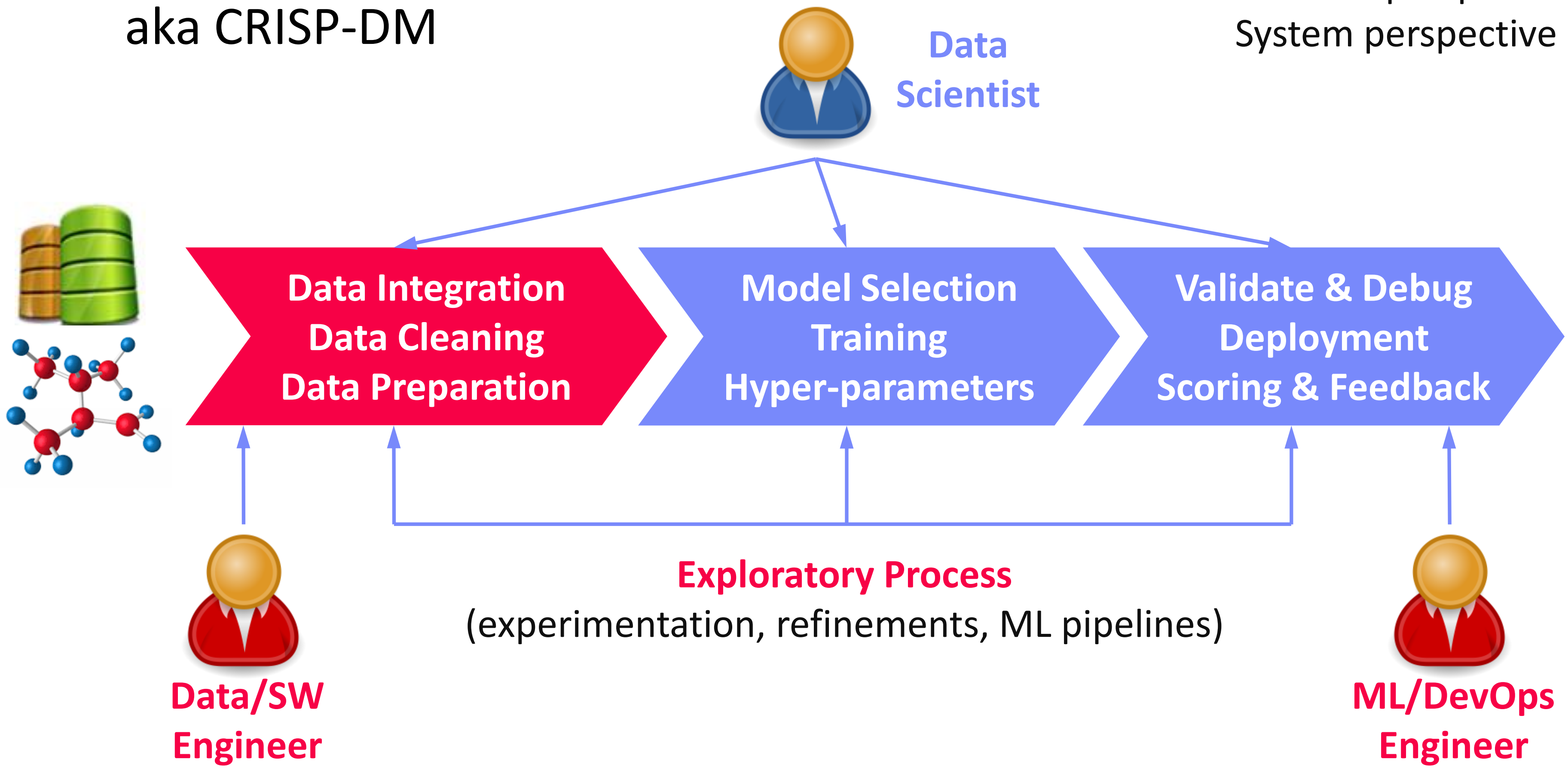


The Data Science Lifecycle

aka KDD Process

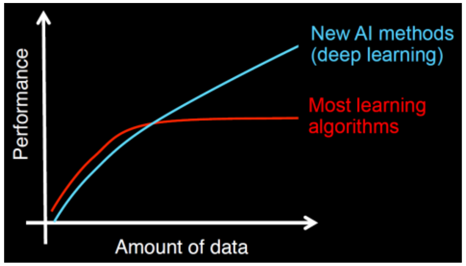
aka CRISP-DM

Data-centric View:
Application perspective
Workload perspective
System perspective



Driving Factors for ML

[**Credit:** Andrew Ng'14]



- **Improved Algorithms and Models**
  - Success across data and application domains (e.g., health care, finance, transport, production)
  - More complex models which leverage large data
- **Availability of Large Data Collections**
  - Increasing automation and monitoring ➔ data (simplified by cloud computing & services)
  - Feedback loops, data programming/augmentation

Feedback Loop

Data → Model → Usage → Data

- **HW & SW Advancements**
  - Higher performance of hardware and infrastructure (cloud)
  - Open-source large-scale computation frameworks, ML systems, and vendor-provides libraries

Stack of ML Systems

Deployment & Scoring

Validation & Debugging

Training

Hyper-parameter Tuning

Model and Feature Selection

Data Programming & Augmentation

Data Preparation (e.g., one-hot, binning)

Data Integration & Data Cleaning

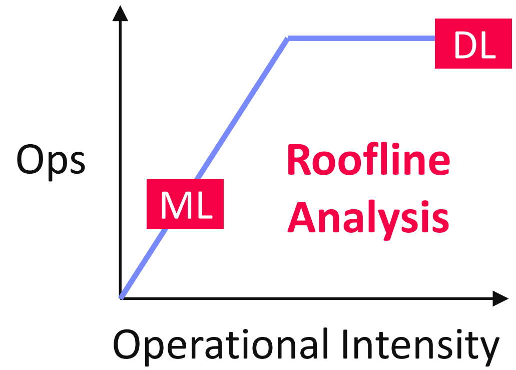
ML Apps & Algorithms	Supervised, unsupervised, RL linear algebra, libs, AutoML
Language Abstractions	Eager interpretation, lazy evaluation, prog. compilation
Fault Tolerance	Approximation, lineage, checkpointing, checksums, ECC
Execution Strategies	Local, distributed, cloud (data, task, parameter server)
Data Representations	Dense & sparse tensor/matrix; compress, partition, cache
HW & Infrastructure	CPUs, NUMA, GPUs, FPGAs, ASICs, RDMA, SSD/NVM

Improve **accuracy** vs. **performance** vs. **resource requirements**

→ **Specialization & Heterogeneity**

Accelerators (GPUs, FPGAs, ASICs)

- **Memory- vs Compute-intensive**
  - **CPU:** dense/sparse, large mem, high mem-bandwidth, moderate compute
  - **GPU:** dense, small mem, slow PCI, very high mem-bandwidth / compute



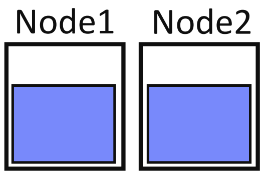
- **Graphics Processing Units (GPUs)**
  - Extensively used for deep learning training and scoring
  - NVIDIA Volta: "tensor cores" for 4x4 mm → 64 2B FMA instruction
- **Field-Programmable Gate Arrays (FPGAs)**
  - Customizable HW accelerators for prefiltering, compression, DL
  - Examples: Microsoft Catapult/Brainwave Neural Processing Units (NPUs)
- **Application-Specific Integrated Circuits (ASIC)**
  - Spectrum of chips: DL accelerators to computer vision
  - Examples: Google TPUs (64K 1B FMA), NVIDIA DLA, Intel NNP

Data Representation

- **ML- vs DL-centric Systems**
  - **ML:** dense and sparse matrices or tensors, different sparse formats (CSR, CSC, COO), frames (heterogeneous)
  - **DL:** mostly dense tensors, embeddings for NLP, graphs

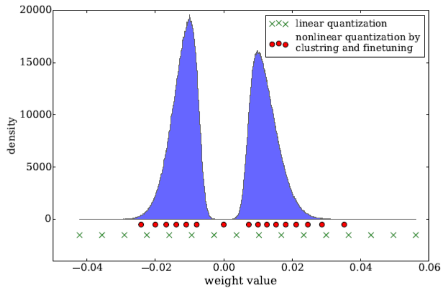
```
vec(Berlin) – vec(Germany)
+ vec(France) ≈ vec(Paris)
```

- **Data-Parallel Operations for ML**
  - Distributed matrices: RDD<MatrixIndexes,MatrixBlock>
  - Data properties: **distributed caching**, **partitioning**, **compression**



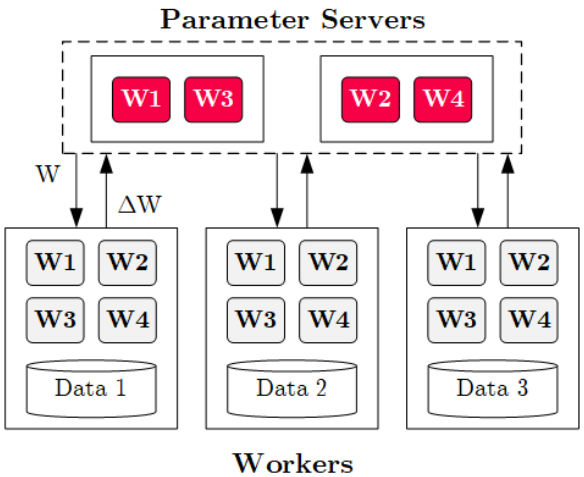
- **Lossy Compression ➔ Acc/Perf-Tradeoff**
  - Sparsification (reduce non-zero values)
  - Quantization (reduce value domain), learned
  - New data types: Intel Flexpoint (mantissa, exp)

[**Credit:** Song Han'16]



Execution Strategies

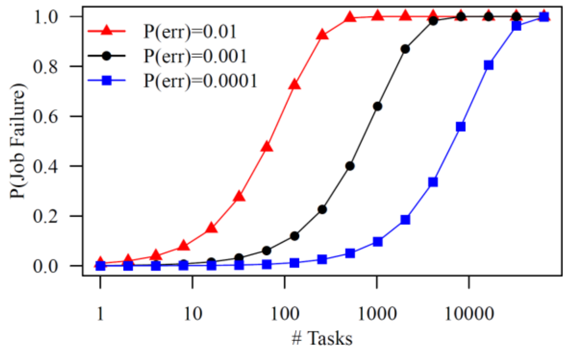
- **Batch Algorithms: Data and Task Parallel**
  - Data-parallel operations
  - Different physical operators
- **Mini-Batch Algorithms: Parameter Server**
  - Data-parallel and model-parallel PS
  - Update strategies (e.g., async, sync, backup)
  - Data partitioning strategies
  - Federated ML (trend 2018)



- **Lots of PS Decisions ➔ Acc/Perf-Tradeoff**
  - Configurations (#workers, batch size/param schedules, update type/freq)
  - Transfer optimizations: lossy compression, sparsification, residual accumulation, layer-wise all-reduce, gradient clipping, momentum corrections

Fault Tolerance & Resilience

- **Resilience Problem**
  - Increasing error rates at scale (soft/hard mem/disk/net errors)
  - Robustness for preemption
  - **Need cost-effective resilience**



- **Fault Tolerance in Large-Scale Computation**
  - Block replication (min=1, max=3) in distributed file systems
  - ECC; checksums for blocks, broadcast, shuffle
  - Checkpointing (MapReduce: all task outputs; Spark/DL: on request)
  - Lineage-based recomputation for recovery in Spark
- **ML-specific Schemes** (exploit app characteristics)
  - Estimate contribution from lost partition to avoid strugglers
  - Example: user-defined "compensation" functions

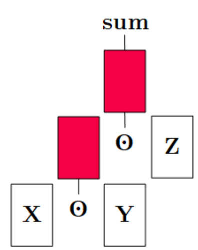
Language Abstractions

- **Optimization Scope**
  - **#1 Eager Interpretation** (debugging, no opt)
  - **#2 Lazy expression evaluation** (some opt, avoid materialization)
  - **#3 Program compilation** (full opt, difficult)

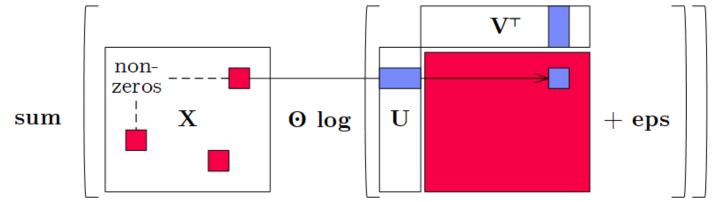
Apache SystemDS

- **Optimization Objective**
  - Most common: **min time** s.t. memory constraints
  - Multi-objective: **min cost** s.t. time, **min time** s.t. acc, **max acc** s.t. time

- **Trend: Fusion and Code Generation**
  - Custom fused operations
  - Examples: SystemDS, Weld, Taco, Julia, TF XLA, TVM, TensorRT



Sparsity-Exploiting Operator



ML Applications

- **ML Algorithms (cost/benefit – time vs acc)**
  - Unsupervised/supervised; batch/mini-batch; first/second-order ML
  - Mini-batch DL: variety of NN architectures and SGD optimizers
- **Specialized Apps: Video Analytics in NoScope (time vs acc)**
  - Difference detectors / specialized models for "short-circuit evaluation"

[**Credit:** Daniel Kang'17]

- **AutoML (time vs acc)**
  - Not algorithms but tasks (e.g., `doClassify(X, y)` + search space)
  - Examples: MLBase, Auto-WEKA, TuPAQ, Auto-sklearn, Auto-WEKA 2.0
  - AutoML services at Microsoft Azure, Amazon AWS, Google Cloud
- **Data Programming and Augmentation (acc?)**
  - Generate noisy labels for pre-training
  - Exploit expert rules, simulation models, rotations/shifting, and labeling IDEs (Software 2.0)

[**Credit:** Jonathan Tremblay'18]

Landscape of ML Systems

Apache SystemML™

#1 Language Abstraction

MAHOUT, julia, PYTORCH, TensorFlow, scikit learn, APACHE Spark, VOWPAL WABBIT, cuDNN, cuBLAS

- **Linear Algebra Programs**
- **Computation Graphs**
- **Algorithm Libraries**
- **Operator Libraries**

#2 Execution Strategies

- **Parameter Server** (Modell-Parallel)
- **Task-Parallel** Constructs
- **Data-Parallel** Operations

TensorFlow, Microsoft CNTK, mxnet, MATLAB, R, APACHE Spark, MAHOUT

#3 Distribution

- **Local** (single node)
- **HW accelerators** (GPUs, FPGAs, ASICs)
- **Distributed**

julia, scikit learn, TensorFlow, Microsoft CNTK, mxnet, APACHE Spark, MAHOUT

#4 Data Types

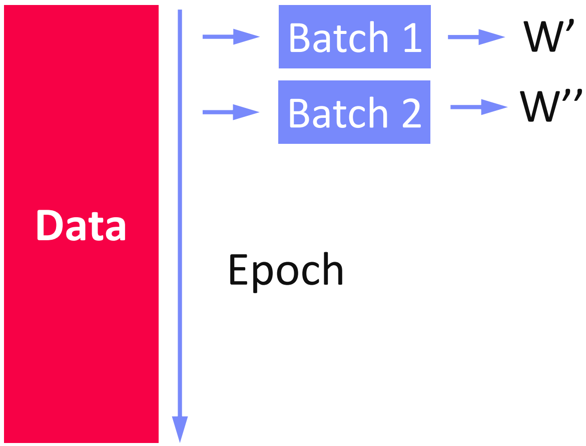
APACHE GIRAPH, turi create intelligence™, APACHE Spark, MAHOUT, julia, TensorFlow, NumPy

- **Collections**
- **Graphs**
- **Matrices**
- **Tensors**
- **Frames**

Distributed Parameter Servers

Background: Mini-batch ML Algorithms

- **Mini-batch ML Algorithms**
  - Iterative ML algorithms, where each iteration only uses a **batch of rows** to make the next model update (in **epochs** or w/ **sampling**)
  - For large and **highly redundant training sets**
  - **Applies to almost all iterative**, model-based ML algorithms (LDA, reg., class., factor., DNN)
  - **Stochastic Gradient Descent** (SGD)



- **Statistical vs Hardware Efficiency** (batch size)
  - **Statistical efficiency:** # accessed data points to achieve certain accuracy
  - **Hardware efficiency:** number of independent computations to achieve high hardware utilization (parallelization at different levels)
  - **Beware higher variance / class skew for too small batches!**

➔ **Training Mini-batch ML algorithms sequentially is hard to scale**

Background: Mini-batch DNN Training (LeNet)

[Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner: Gradient-Based Learning Applied to Document Recognition, **Proc of the IEEE 1998**]

```
# Initialize W1-W4, b1-b4
# Initialize SGD w/ Nesterov momentum optimizer
iters = ceil(N / batch_size)

for( e in 1:epochs ) {
   for( i in 1:iters ) {
      X_batch = X[((i-1) * batch_size) %% N + 1:min(N, beg + batch_size - 1),]
      y_batch = Y[((i-1) * batch_size) %% N + 1:min(N, beg + batch_size - 1),]

      ## layer 1: conv1 -> relu1 -> pool1
      ## layer 2: conv2 -> relu2 -> pool2
      ## layer 3:  affine3 -> relu3 -> dropout
      ## layer 4:  affine4 -> softmax
      outa4 = affine::forward(outd3, W4, b4)
      probs = softmax::forward(outa4)

      ## layer 4:  affine4 <- softmax
      douta4 = softmax::backward(dprobs, outa4)
      [doutd3, dW4, db4] = affine::backward(douta4, outr3, W4, b4)
      ## layer 3:  affine3 <- relu3 <- dropout
      ## layer 2: conv2 <- relu2 <- pool2
      ## layer 1: conv1 <- relu1 <- pool1

      # Optimize with SGD w/ Nesterov momentum W1-W4, b1-b4
      [W4, vW4] = sgd_nesterov::update(W4, dW4, lr, mu, vW4)
      [b4, vb4] = sgd_nesterov::update(b4, db4, lr, mu, vb4)
   }
}
```

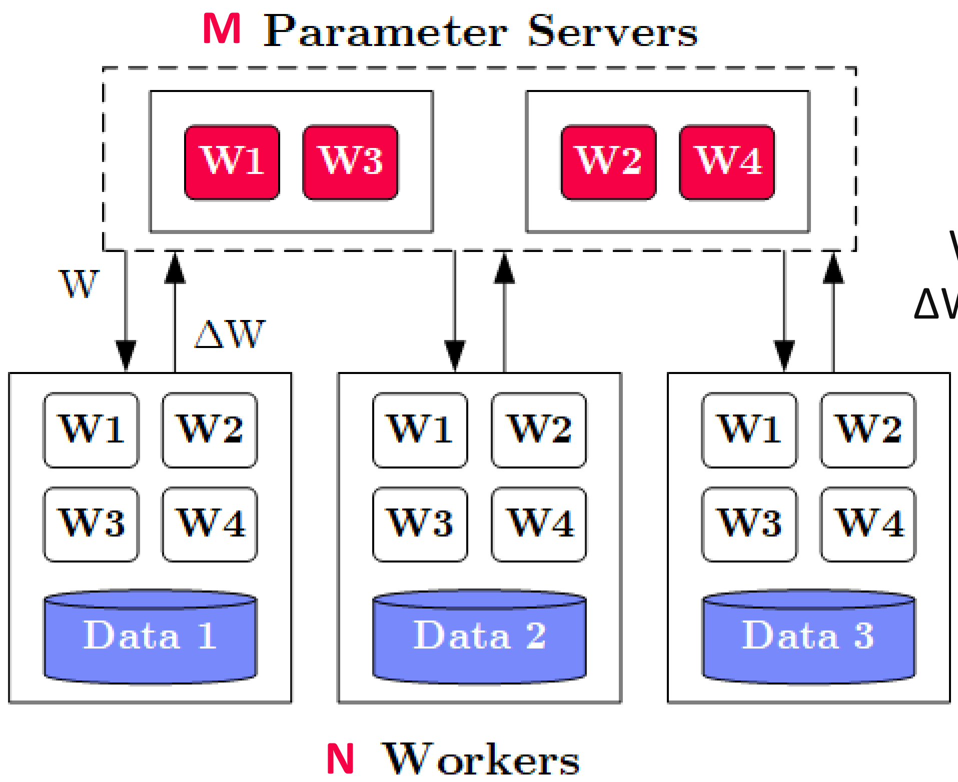
NN Forward Pass

NN Backward Pass
→ Gradients

Model Updates

Overview Data-Parallel Parameter Servers

- **System Architecture**
  - M Parameter Servers
  - N Workers
  - Optional Coordinator



W .. Model
ΔW .. Gradient

- **Key Techniques**
  - Data partitioning D → workers Di (e.g., disjoint, reshuffling)
  - Updated strategies (e.g., synchronous, asynchronous)
  - Batch size strategies (small/large batches, hybrid methods)

History of Parameter Servers

- **1st Gen: Key/Value**
  - **Distributed key-value store** for parameter exchange and synchronization
  - Relatively high overhead
- **2nd Gen: Classic Parameter Servers**
  - **Parameters as dense/sparse matrices**
  - Different **update/consistency strategies**
  - Flexible configuration and fault tolerance
- **3rd Gen: Parameter Servers w/ improved data communication**
  - Prefetching and range-based pull/push
  - Lossy or lossless compression w/ compensations
- **Examples**
  - TensorFlow, MXNet, PyTorch, CNTK, Petuum

[Alexander J. Smola, Shravan M. Narayanamurthy: An Architecture for Parallel Topic Models. **PVLDB 2010**]

[Jeffrey Dean et al.: Large Scale Distributed Deep Networks. **NIPS 2012**]

[Mu Li et al: Scaling Distributed Machine Learning with the Parameter Server. **OSDI 2014**]

[Jiawei Jiang, Bin Cui, Ce Zhang, Lele Yu: Heterogeneity-aware Distributed Parameter Servers. **SIGMOD 2017**]

[Jiawei Jiang et al: SketchML: Accelerating Distributed Machine Learning with Data Sketches. **SIGMOD 2018**]

Basic Worker Algorithm (batch)

```
for( i in 1:epochs ) {
   for( j in 1:iterations ) {
      params = pullModel(); # W1-W4, b1-b4 lr, mu
      batch = getNextMiniBatch(data, j);
      gradient = computeGradient(batch, params);
      pushGradients(gradient);
   }
}
```

[Jeffrey Dean et al.: Large Scale Distributed Deep Networks. **NIPS 2012**]

Extended Worker Algorithm (nfetch batches)

```
gradientAcc = matrix(0,...);
for( i in 1:epochs ) {
   for( j in 1:iterations ) {
      if( step mod nfetch = 0 )
         params = pullModel();
      batch = getNextMiniBatch(data, j);
      gradient = computeGradient(batch, params);
      gradientAcc += gradient;
      params = updateModel(params, gradients);
      if( step mod nfetch = 0 ) {
         pushGradients(gradientAcc); step = 0;
         gradientAcc = matrix(0, ...);
      }
      step++;
}  }
```

nfetch batches require **local gradient accrual** and **local model update**

[Jeffrey Dean et al.: Large Scale Distributed Deep Networks. **NIPS 2012**]

Update Strategies

- **Bulk Synchronous Parallel (BSP)**
  - Update model w/ accrued gradients
  - Barrier for N workers

- **Asynchronous Parallel (ASP)**
  - Update model for each gradient
  - No barrier

but, stale model updates

- **Synchronous w/ Backup Workers**
  - Update model w/ accrued gradients
  - Barrier for N of N+b workers

[Martín Abadi et al: TensorFlow: A System for Large-Scale Machine Learning. **OSDI 2016**]

Intro to LLMs

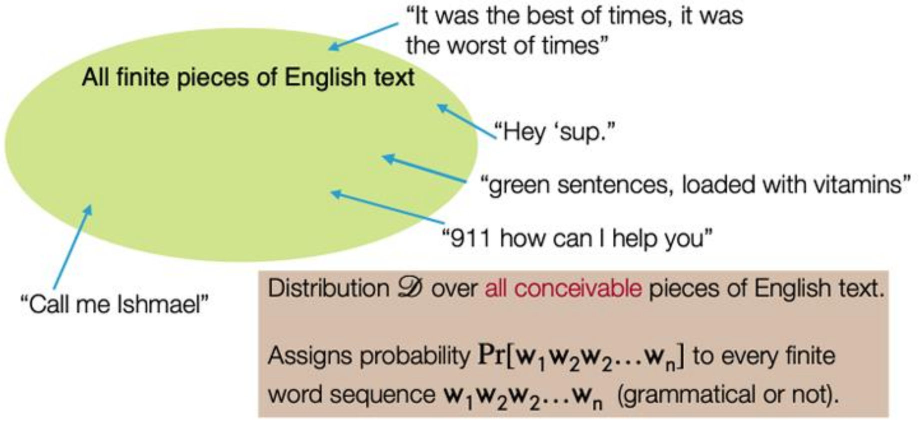
- Google Gemini
- ChatGPT
- Lamda
- Llama
- Gork
- Mistral
- Eliza (1966)

	Gemini Ultra	Gemini Pro	GPT-4	GPT-3.5	PaLM 2-L	Claude 2	Inflect-ion-2	Grok 1	LLAMA-2
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	**90.04%** CoT@32* 83.7% 5-shot	79.13% CoT@8* 71.8% 5-shot	87.29% CoT@32 (via API**) 86.4% 5-shot (reported)	70% 5-shot	78.4% 5-shot	78.5% 5-shot CoT	79.6% 5-shot	73.0% 5-shot	68.0%***
GSM8K Grade-school math (Cobbe et al., 2021)	**94.4%** Maj1@32	86.5% Maj1@32	92.0% SFT & 5-shot CoT	57.1% 5-shot	80.0% 5-shot	88.0% 0-shot	81.4% 8-shot	62.9% 8-shot	56.8% 5-shot
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	**53.2%** 4-shot	32.6% 4-shot	52.9% 4-shot (via API**) 50.3% (Zheng et al., 2023)	34.1% 4-shot (via API**)	34.4% 4-shot	—	34.8%	23.9% 4-shot	13.5% 4-shot
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	**83.6%** 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)	66.6% 3-shot (via API**)	77.7% 3-shot	—	—	—	51.2% 3-shot
HumanEval Python coding tasks (Chen et al., 2021)	**74.4%** 0-shot (IT)	67.7% 0-shot (IT)	67.0% 0-shot (reported)	48.1% 0-shot	—	70.0% 0-shot	44.5% 0-shot	63.2% 0-shot	29.9% 0-shot
Natural2Code Python code generation. (New held-out set with no leakage on web)	**74.9%** 0-shot	69.6% 0-shot	73.9% 0-shot (via API**)	62.3% 0-shot (via API**)	—	—	—	—	—
DROP Reading comprehension & arithmetic. (metric: F1-score) (Dua et al., 2019)	**82.4** Variable shots	74.1 Variable shots	80.9 3-shot (reported)	64.1 3-shot	82.0 Variable shots	—	—	—	—
HellaSwag (validation set) Common-sense multiple choice questions (Zellers et al., 2019)	87.8% 10-shot	84.7% 10-shot	**95.3%** 10-shot (reported)	85.5% 10-shot	86.8% 10-shot	—	89.0% 10-shot	—	80.0%***
WMT23 Machine translation (metric: BLEURT) (Tom et al., 2023)	**74.4** 1-shot (IT)	71.7 1-shot	73.8 1-shot (via API**)	—	72.7 1-shot	—	—	—	—

https://blog.google/technology/ai/google-gemini-ai/#sundar-note

Intro to LLMs

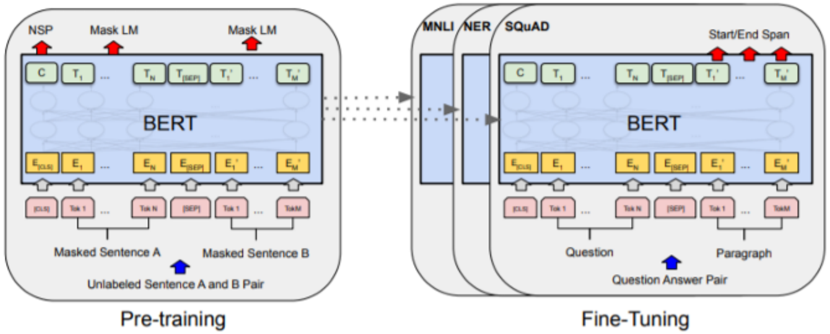
- Next work prediction
- A probabilistic model that assign probability to every finite sequence in English language
- considering context, position, grammar and structure
- Sentence “the cat set on the mat”
  - P(the cat sat on the map) = P(the)*P(cat|the) * P(sat|the cat) *P(on |the cat sat)*P(the|the cat sat on) *P(mat|the cat sat on the)

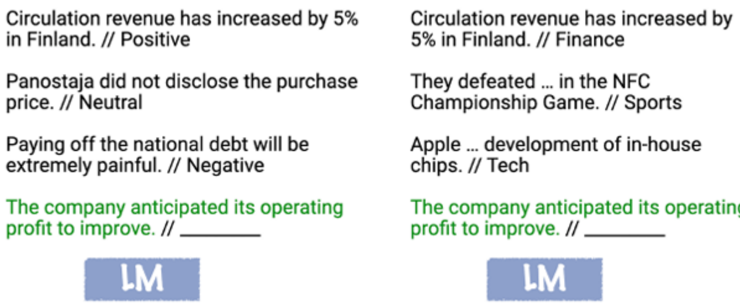


Source: COS 324

LLM Training

- **Transformer based neural networks**
- **Pre-training (expensive)**
  - Download ~10TB of text.
  - Get a cluster of ~6,000 GPUs.
  - Compress the text into a neural network, pay ~$2M, wait ~12 days.
  - Obtain base model.
- **Fine Tuning**
  - Write labeling instructions
  - Hire people (or use scale.ai!), collect 100K high quality ideal Q&A responses, and/or comparisons.
  - Finetune base model on this data → ~1 day.
  - Obtain assistant model.
  - Run a lot of evaluations.





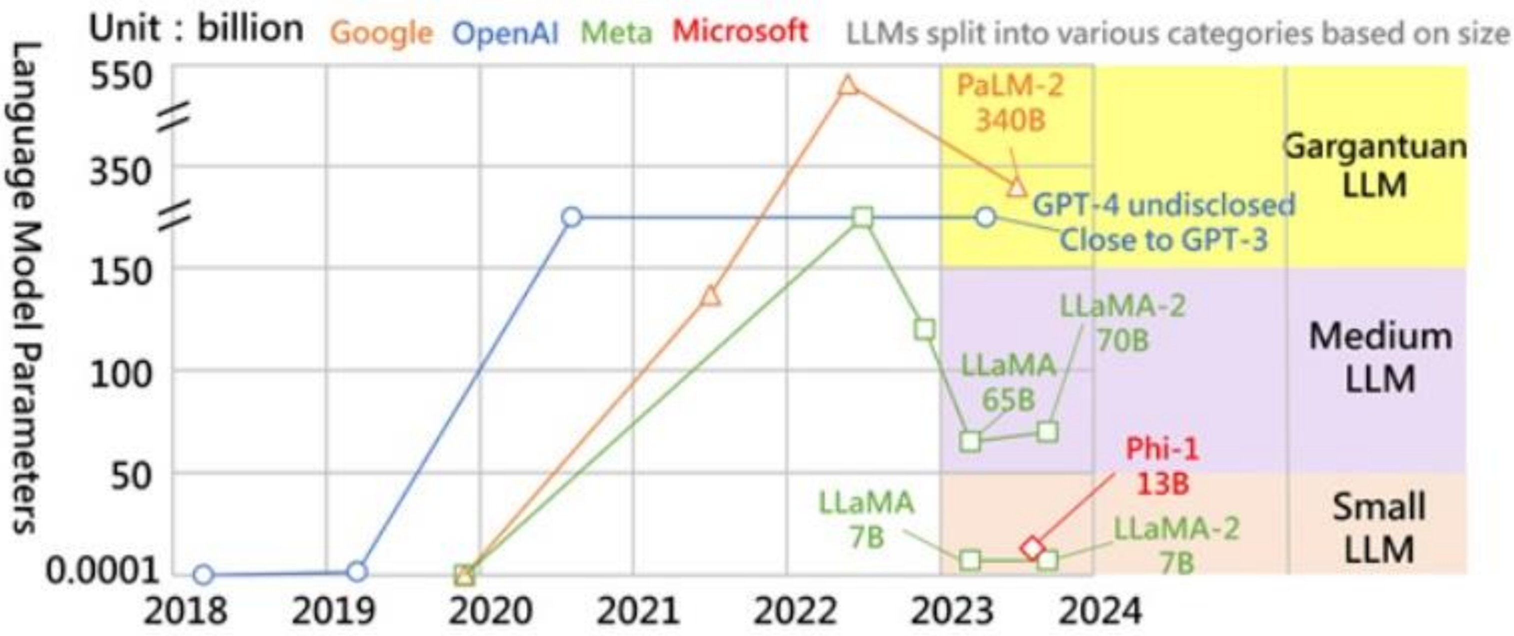
https://ai.stanford.edu/blog/understanding-incontext/

https://www.youtube.com/watch?v=zjkBMFhNj_g

LLM Parameters

[Ashish Vaswani et. al., Attention is All you Need. **NeurIPS 2017**]

- **Transformer based neural networks**



Source: Compiled by DIGITIMES Research, Aug. 2023
https://www.digitimes.com/news/a20231221VL202/2024-outlook-ai-edge-ai-llm.html

Q&A and Exam Preparation

Selected Example Questions

Multiple choice question (40/100)

- **Given the dataset identify the type of missingness.**
  a. Missing Completely at Random (MCAR)
  b. Missing at Random (MAR)
  c. Missing Not at Random (MNAR)
  d. All of above

Profession	Salary
A	2000
C	5000
B	2300
C	?
C	?
A	2000

Data Warehousing

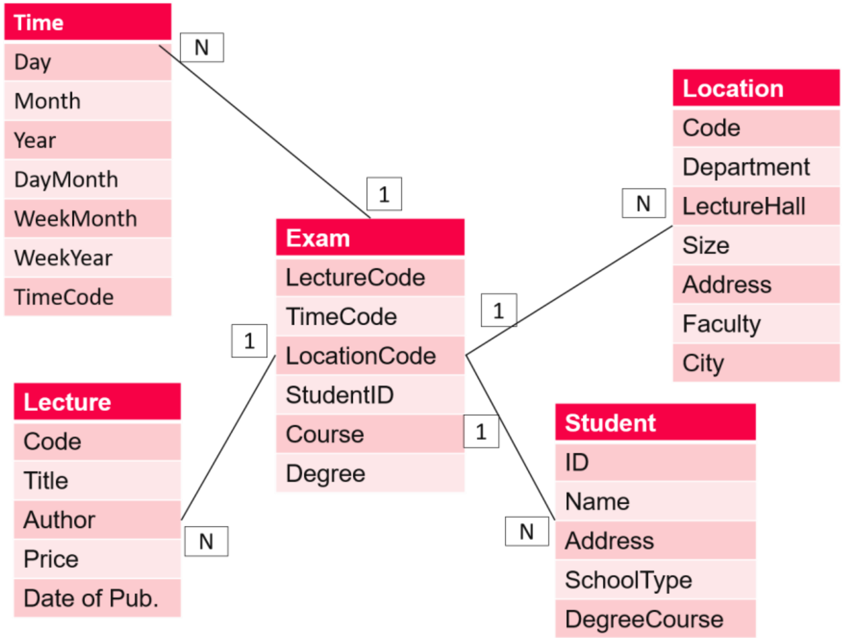
- **Create a** star schema **from the given relational schema. Identify the fact and dimension tables and indicate the relationship type (cardinality)** [10/100 points].

EXAM (TimeCode, LocationCode, Student, LectureCode, Course, Degree)

TIME(TimeCode, Day, Month, Year, WeekMonth, WeekYear, DayMonth)

LOCATION(LocationCode,LectureHall, Department, Faculty, Address, City)

STUDENT (ID, Name, Address, DegreeCourse)



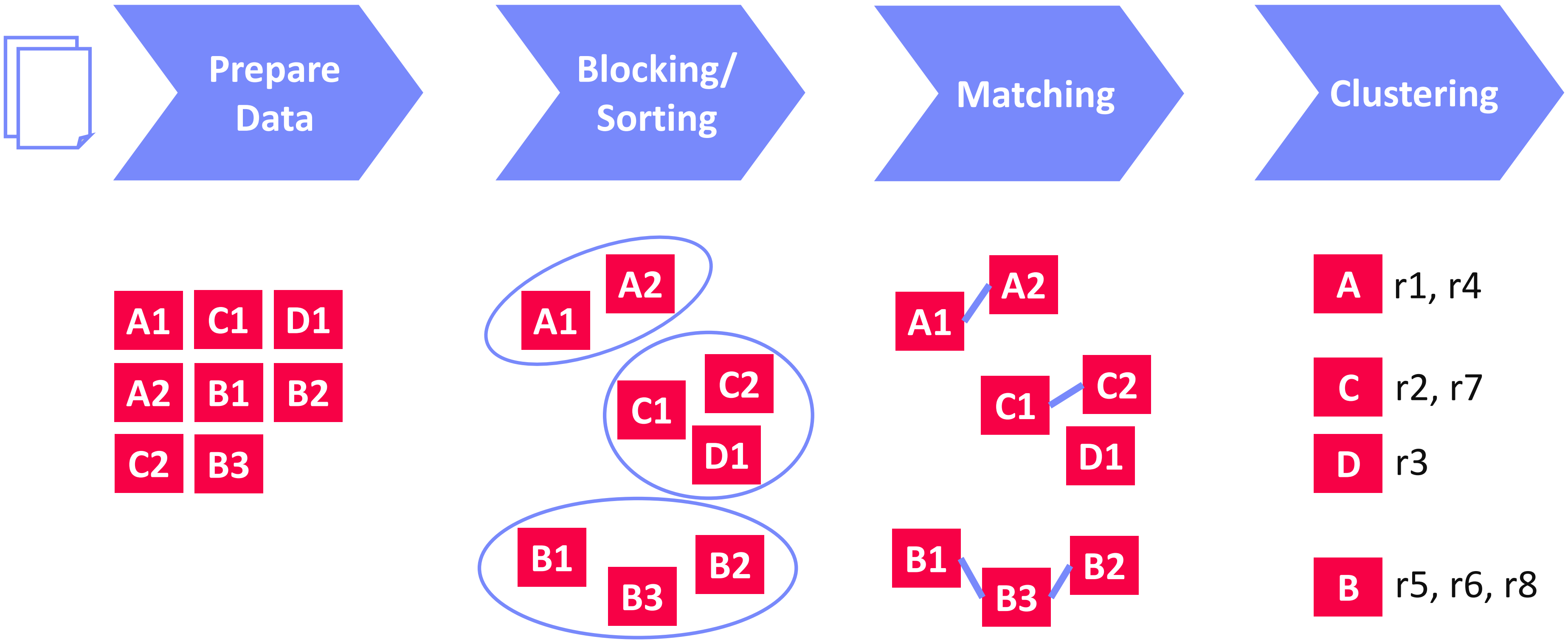
Message-oriented Middleware

- **Describe the Message Delivery Guarantees At-Most-Once, At-Least-Once and Exactly-Once, and indicate which of them require persistent storage before sending.** [6/100 points]

Name	Description	Storage
At-Most-Once	Send and forget, never sent message twice (even on failures)	No
At-Least-Once	Store and forward, replay stream from (acknowledged) checkpoint	Yes
Exactly-Once	Store and forward, replay stream from (acknowledged) checkpoint, transactional delivery	Yes

Schema Matching / Entity Linking

- **Explain the phases of a typical Entity Resolution Pipeline with example techniques for the individual phases.** [20/100 points]



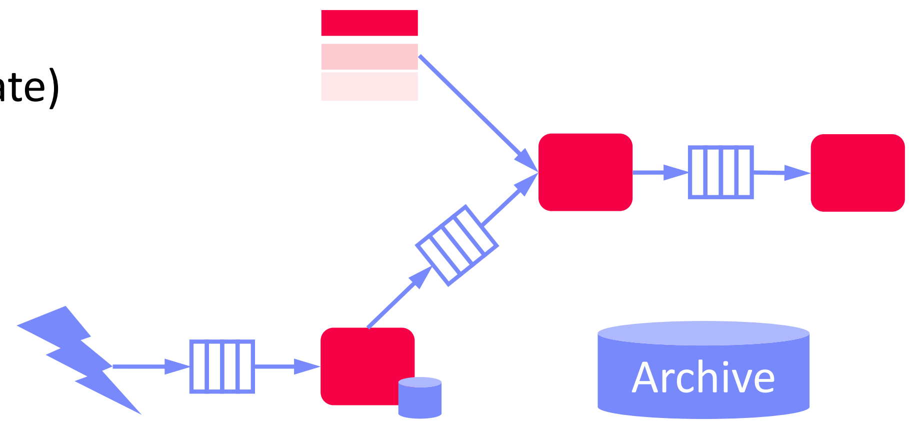
Data Parallel Computation / Stream Mining

- **Assume three nodes with CPU and memory capacity N1 (32 cores, 64 GB), N2 (16 cores, 64 GB), N3 (64 cores, 128 GB) and a stream of resource requests R1 . . . R7. Schedule these requests to available resources (assign requests to nodes) in order to maximize the number of fulfilled requests. (4 points)**
- **R1: (30 cores, 8 GB)**
- **R2: (6 cores, 32 GB)**
- **R3: (8 cores, 64 GB)**
- **R4: (10 cores, 32 GB)**
- **R5: (8 cores, 32 GB)**
- **R6: (16 cores, 64 GB)**
- **R7: (16 cores, 16 GB)**

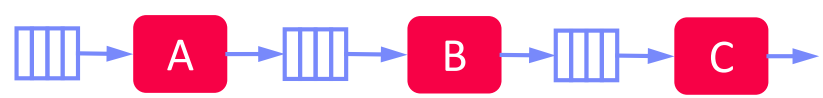
	N1 (32/64)	N2 (16/64)	N3 (64/128)
requests	R6	R2, R4	R1, R3, R5, R7
utilization	16/64	16/64	62/120

Stream Processing

- **Describe the concept of Continuous Queries and a Basic System Infrastructure to process them over incoming data streams.** [8/100 points]
  - Deployed Data flow graphs
  - **Nodes:** asynchronous ops (w/ state) (e.g., separate threads / queues)
  - **Edges:** data dependencies (tuple/message streams)
  - **Push model:** data production controlled by source



- **Given the continuous query A-B-C, what are the resulting perf characteristics?** [4 points]
  - Max throughput
  - Min tuple latency



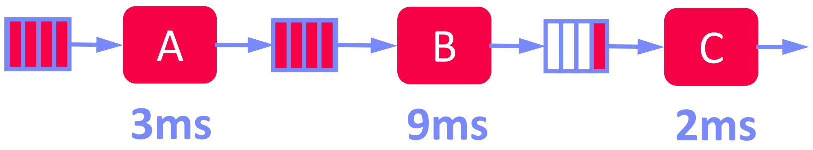
$1/\max(C(op_i)) = 1/5 \text{ tuples/ms} = 200 \text{ tuples/s}$

$\text{sum}(C(op_i)) = 1\text{ms} + 5\text{ms} + 2\text{ms} = 8\text{ms}$

Stream Processing, cont.

- **Describe the three classes of techniques for handling overload situations in continuous queries?** [6/100 points]

- **#1 Back Pressure**
  - Graceful handling of overload w/o data loss
  - Slow down sources
  - E.g., blocking queues



Self-adjusting operator scheduling
Pipeline runs at rate of slowest op

- **#2 Load Shedding**
  - #1 Random-sampling-based load shedding
  - #2 Relevance-based load shedding
  - #3 Summary-based load shedding (synopses)
- **#3 Distributed Stream Processing** (see last part)
  - Data flow partitioning (distribute the query)
  - Key range partitioning (distribute the data stream

Summary and Q&A

- **Landscape of ML Systems**
- **Distributed Parameter Servers**
- **Large Language Models**
- **Q&A and Exam Preparation**

Thanks

(please, participate in the **course evaluation**)

- **Oral Exam**
  - Starting **Jan 22**
- **Exam**
  - Exam date: **Feb 02**