“If we have data, let’s look at data.
If all we have are opinions, let’s go with mine.”

Jim Barksdale, former Netscape CEO

Lecture:
Graph Parallel Processing

CS205: Computing Foundations for Computational Science

Bill Richmond
AI/ML Evangelist
Amazon Web Services

Spring Term 2020

INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

HARVARD
School of Engineering and Applied Sciences

Before We Start

Where We Are

Computing Foundations for Computational and Data Science

How to use modern computing platforms in solving scientific problems

Intro: Large-Scale Computational and Data Science

A. Parallel Processing Fundamentals

B. Parallel Computing

C. Parallel Data Processing

- C.1. Batch Data Processing
- C.2. Dataflow Processing
- C.3. Stream Data Processing
  - Graph Parallel Processing

Wrap-Up: Advanced Topics

Next Steps

- HWC due on Monday 4/20!

- Final Project (next milestones):
  - Team formation and tentative topic: 3/30
  - Project proposal (4/14 and 4/16)
  - Project design (4/21 and 4/23)
  - Project presentation (5/11)
  - More info at:
    - https://harvard-iacs.github.io/2020-CS205/

Project Requirements

- Demonstrate the need for big compute and/or big data processing, and what can be achieved thanks to large-scale parallel processing.
- Solve a problem for a non-trivial computation graph and with hierarchical parallelism.
- Be implemented on a distributed-memory architecture with either a many-core or a multi-core compute node, and evaluated on at least 8 compute nodes (note: each compute node on Cannon is a multi-core with 32, or 64 cores or with a many-core GPU with hundreds of cores)
- Use a hybrid parallel program in either, for example: MPI + OpenMP, MPI + OpenACC (or OpenCL), Spark or MapReduce + OpenACC (or OpenCL) or MPI + Spark or MapReduce
- Be evaluated on large data sets or problem sizes to demonstrate both weak and strong scaling using appropriate metrics (throughput, efficiency, iso-efficiency...).

Project Proposal Presentation

You will have **5, and ONLY 5, minutes** to briefly summarize your proposal answering bellow questions. You have to prepare 4 slides for your proposal. We will enforce the 5-minute time limit.

What is the **problem** you are trying to solve with this application?

What is the **need for big compute and/or big data processing** and what can be achieved thanks to large-scale parallel processing?

Describe your **model and/or data** in detail: where does it come from, what does it mean, etc.

Which **tools and infrastructures** you are planning to use to build the application?

Zoom Presentation Guidelines

Appoint 1 group member to share their screen on Zoom.

Each group member should present.

Practice ahead of time!

Make sure your mics are muted when you are not presenting.

Don’t forget to stop sharing your screen when your group is done!

CS205: Contents

APPLICATION SOFTWARE

APPLICATION PARALLELISM

PARALLEL PROGRAM DESIGN

PROGRAMMING MODEL

Optimization

OpenACC

OpenMP

MPI

Spark

Map-Reduce

B. BIG COMPUTE

PLATFORM

C. BIG DATA

amazon web services

Open Nebula

FAS RC

ODYSSEY HARVARD FAS RESEARCH COMPUTING

FAS RC

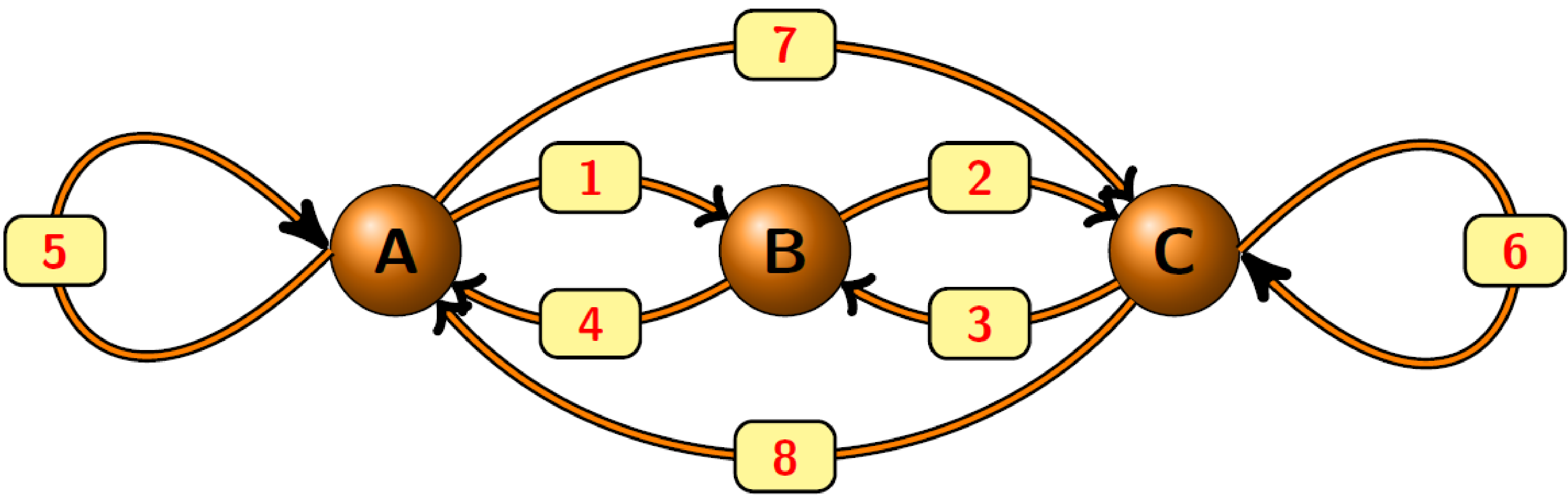
CLOUD COMPUTING

PARALLEL ARCHITECTURES

Review of Graph Theory – What is a Graph?

Graphs

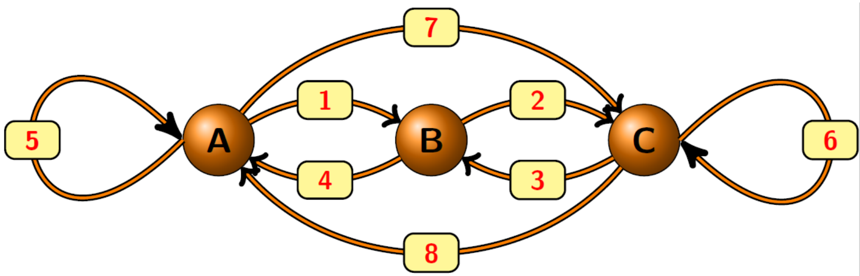
A graph is a mathematical structure that helps us visualize and analyze problems.



Review of Graph Theory – Basic Graph Anatomy

A Definition of a Graph

A graph G consists of a finite, nonempty set of objects called vertices V and a set of 2-element subsets of V called edges E. A graph is often denoted by $G = (V, E)$.



Vertices The "points" in the graph

Edges The "lines" connecting the vertices

Degree The number of vertices

Size The number of edges

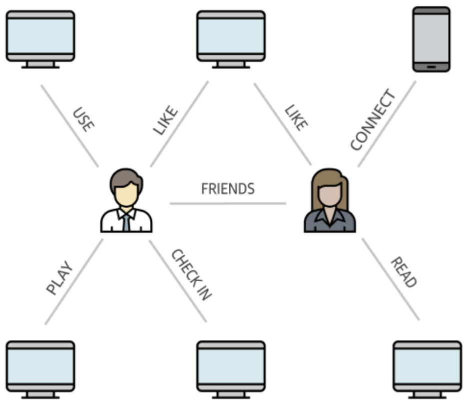
Labeled vs. Unlabeled Vertices can be labeled or not

Multigraph Two vertices can be connected by more than one edge

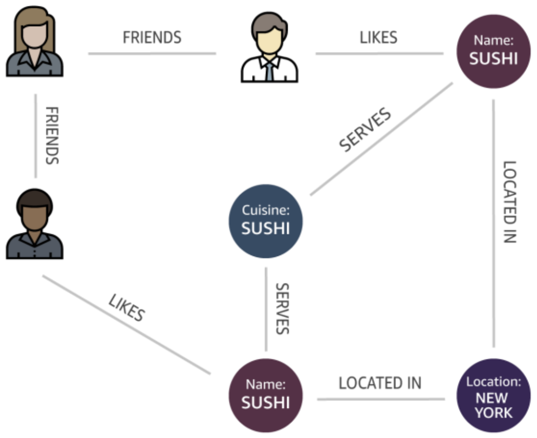
Weighted Edges are labeled with weights

Directed graph An edge between two vertices is directed from one vertex to another, but not the other way around

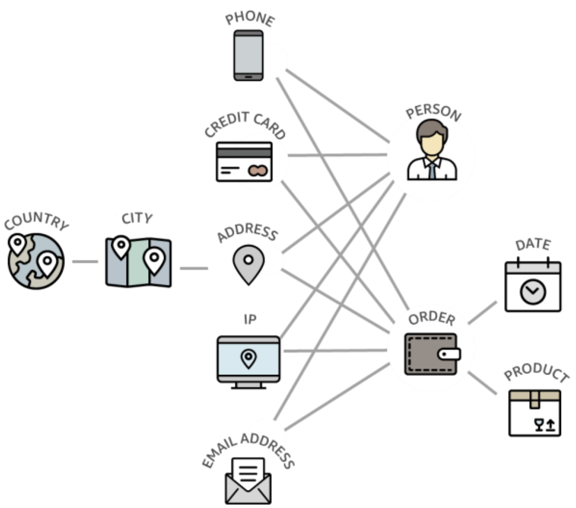
Highly Connected Data



Social Networks



Restaurant Recommendations



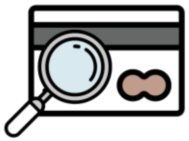
Retail Fraud Detection

Use cases for highly connected data

Social Networking

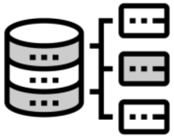
Recommendations

Knowledge Graphs



Fraud Detection

Life
Sciences

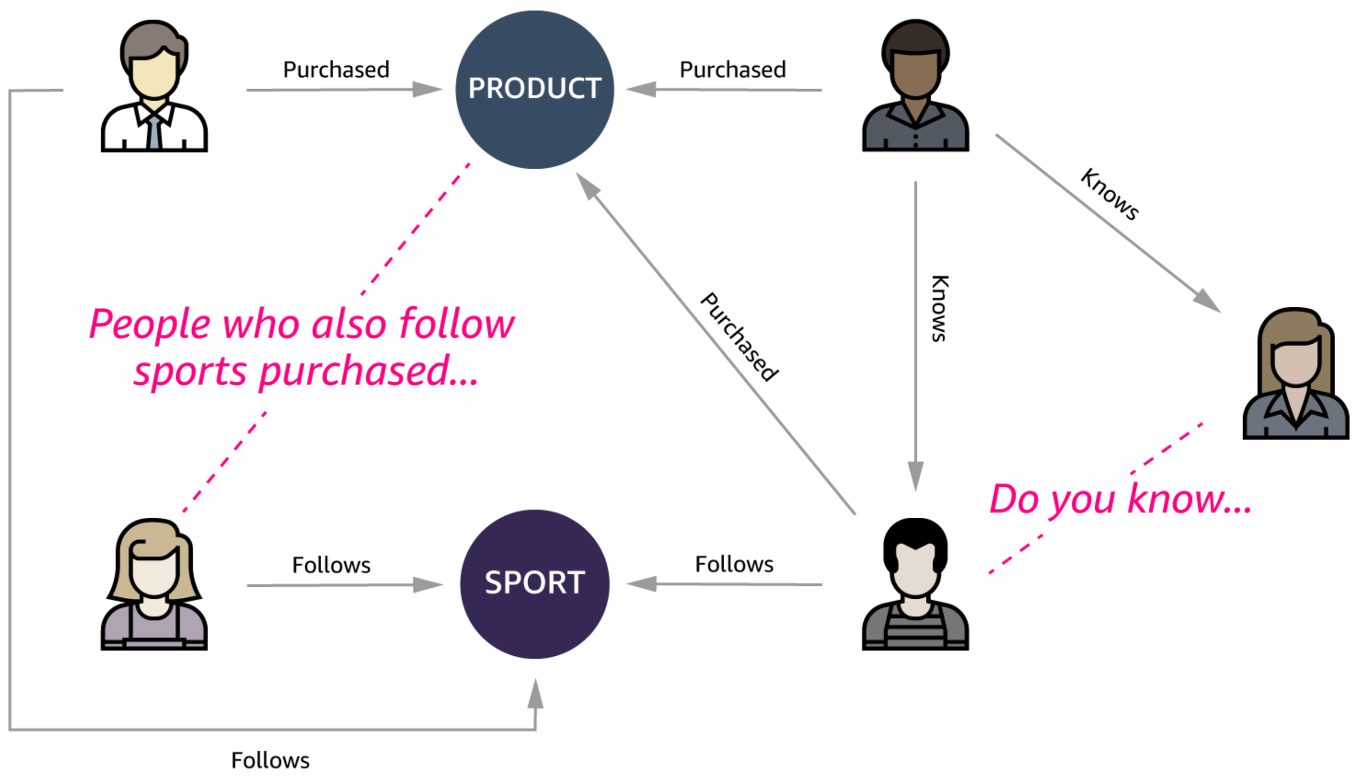


Network & IT Operations

Examples of connected data queries

- Which friends and colleagues do we have in common?
- Which applications and services in my network will be affected if a particular network element – a router or switch, for example – fails? Do we have redundancy throughout the network for our most important customers?
- What's the quickest route between two stations on the underground?
- What do you recommend this customer should buy, view, or listen to next?
- Which products, services and subscriptions does a user have permission to access and modify?
- What's the cheapest or fastest means of delivering this parcel from A to B?
- Which parties are likely working together to defraud their bank or insurer?
- Which institutions are most at risk of poisoning the financial markets?

Recommendations based on relationships

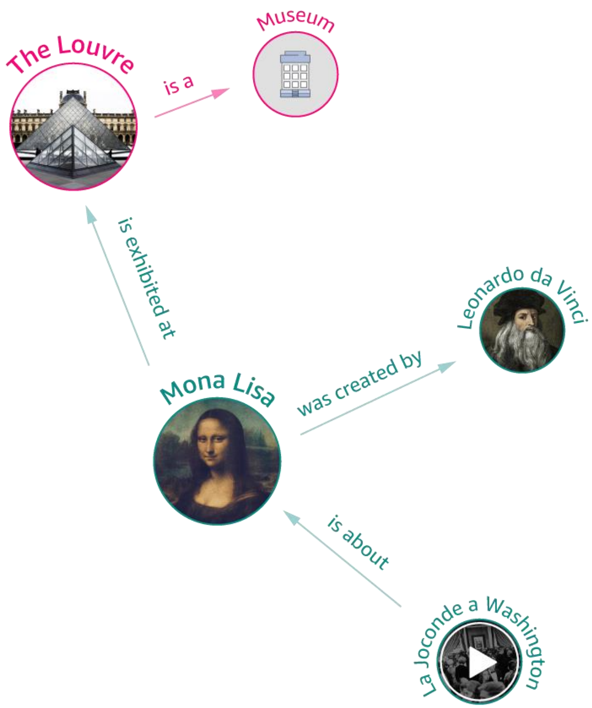


Knowledge Graph Applications

Who painted the
Mona Lisa?

What museums should
Alice visit while in Paris?

What artists have
paintings in The Louvre?



The need for Graph Parallel Processing

- Graphs representing real-world phenomena can be very large
- Development of parallel algorithms for processing large datasets is a very active area of research
- Parallel processing of large graphs

How large is large?

Example: Facebook

- 2.50B+ Monthly Active Users
- 1.66B+ Daily Active Users
- 1.15B+ Mobile Daily Active Users
- 94% of advertising revenue comes from mobile ads
- 83 million fake profiles
- Like & Share Buttons are viewed across 10M websites daily
- Age 25 to 34 (30% of users) is the most common age demographic
- 50% of 18-24 year-olds go on Facebook when they wake up
- Highest traffic occurs mid-week between 1 to 3 pm
- A 7pm post will result in more clicks on average than posting at 8pm
- On Thursdays and Fridays, engagement is 18% higher
- One in five page views in the United States occurs on Facebook
- Every 60 seconds on Facebook: 510K comments are posted, 293K statuses are updated, and 136K photos are uploaded

Imagine trying to use all of this related data (and much more – who is friends with who, is interested in what, etc.) to effectively run a business!

Source: Zephoria

Example: Siemens Smart Infrastructure

- In 2018, there were eight billion devices connected to the internet; by 2030, there will be about one trillion, according to a report by the World Economic Forum. These connected devices include components of the systems that buildings use for vital functions like fire prevention, security and access, HVAC, lighting, and power. A typical smart office building has about 60 types of sensors generating more than 500 MB of data a day—a volume projected to double every two years.

- Siemens Smart Infrastructure focuses on connecting energy systems, buildings, and industry and is interested in the prospect of increasingly sophisticated sensors generating more and more building data. For example, if an HVAC system could use an access-control system's data, it could automatically increase the air conditioning as a conference room fills up, and then turn it down again after the meeting is over.

- Siemens uses graph databases (Amazon Neptune) to model the complex object dependencies in building-generated datasets.

Example: FINRA

- The Financial Industry Regulatory Authority (FINRA) writes and enforces rules governing the activities of more than 3,800 broker-dealers representing more than 600,000 brokers, examines firms for compliance, fosters market transparency, and educates investors.

- Every day, FINRA oversees up to 75 billion market events—99 percent of equities trades and 65 percent of options trades in the United States—applying data analytics to uncover insider trading and other strategies used to gain an unfair advantage.

- Broker-dealers must submit daily electronic data to FINRA, adding up to more than 50,000 files. As soon as data is received, FINRA validates it to ensure it is complete and correctly formatted according to a set of more than 200 rules. The system performs up to half a trillion validations each day. Processing demand varies significantly over time and can double or triple in response to market conditions that drive higher trading volumes.

Some Graph Algorithms

Traversal Visit vertices of a graph in a certain order

- Depth-First Search and Breadth-First Search are examples

Topological Sort Sort the vertices according to some criterion

Strongly Connected Components Maximal strongly connected subgraphs of the graph *G*

- A graph is strongly connected if all vertices are reachable from every vertex.

Label Propagation Determine communities in a network

PageRank Rank the relative importance of vertices in a graph

Triangle Count Counts the number of triangles passing through a given vertex

GraphX vs GraphFrames

GraphX

- GraphX is part of Spark
- Extends the Spark RDD by introducing a Graph abstraction
- Provides a suite of graph operations, builders, and algorithms
- Only works with Scala

GraphFrames

- Not part of Spark
- But, it is designed to be used with Spark
- Graphs are built on Spark `dataframes` rather than RDDs
- Contains the same functionality as GraphX and more
- Works with Scala, Java, and Python

Review of DataFrames

Python DataFrames in Pandas

A 2D labeled data structure with columns of potentially different types

```
import pandas as pd
```

```
d = {'one' : [1., 2., 3., 4.], 'two' : [4., 3., 2., 1.]}
```

```
pd.DataFrame(d)
```

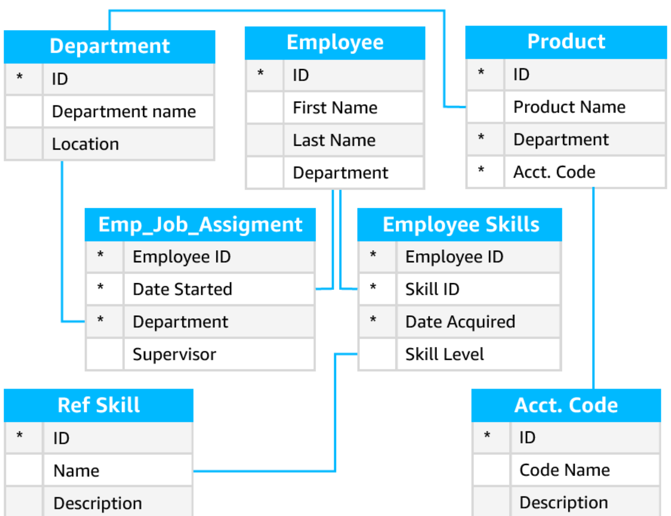
	one	two
0	1.0	4.0
1	2.0	3.0
2	3.0	2.0
3	4.0	1.0

Spark DataFrames

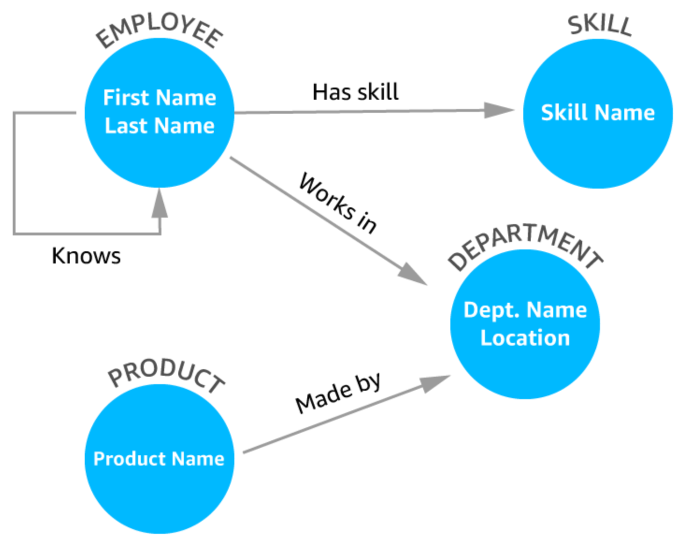
Conceptually equivalent to Python DataFrames, but contain some richer optimizations

And now,
to go a little off-topic…

Different approaches for highly connected data



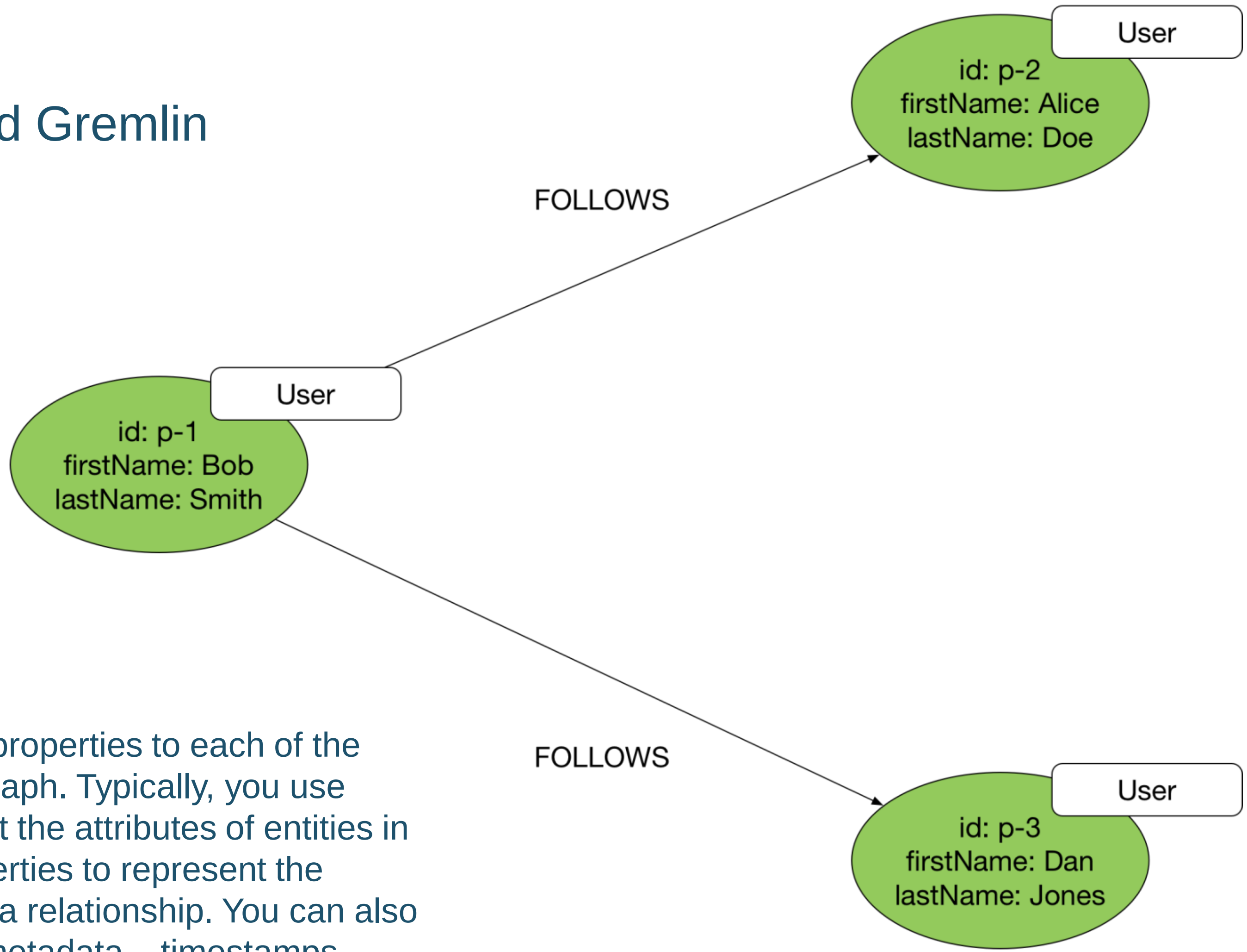
Purpose-built for a business process



Purpose-built to answer questions about relationships

Different data models and query languages

Property Graph and Gremlin



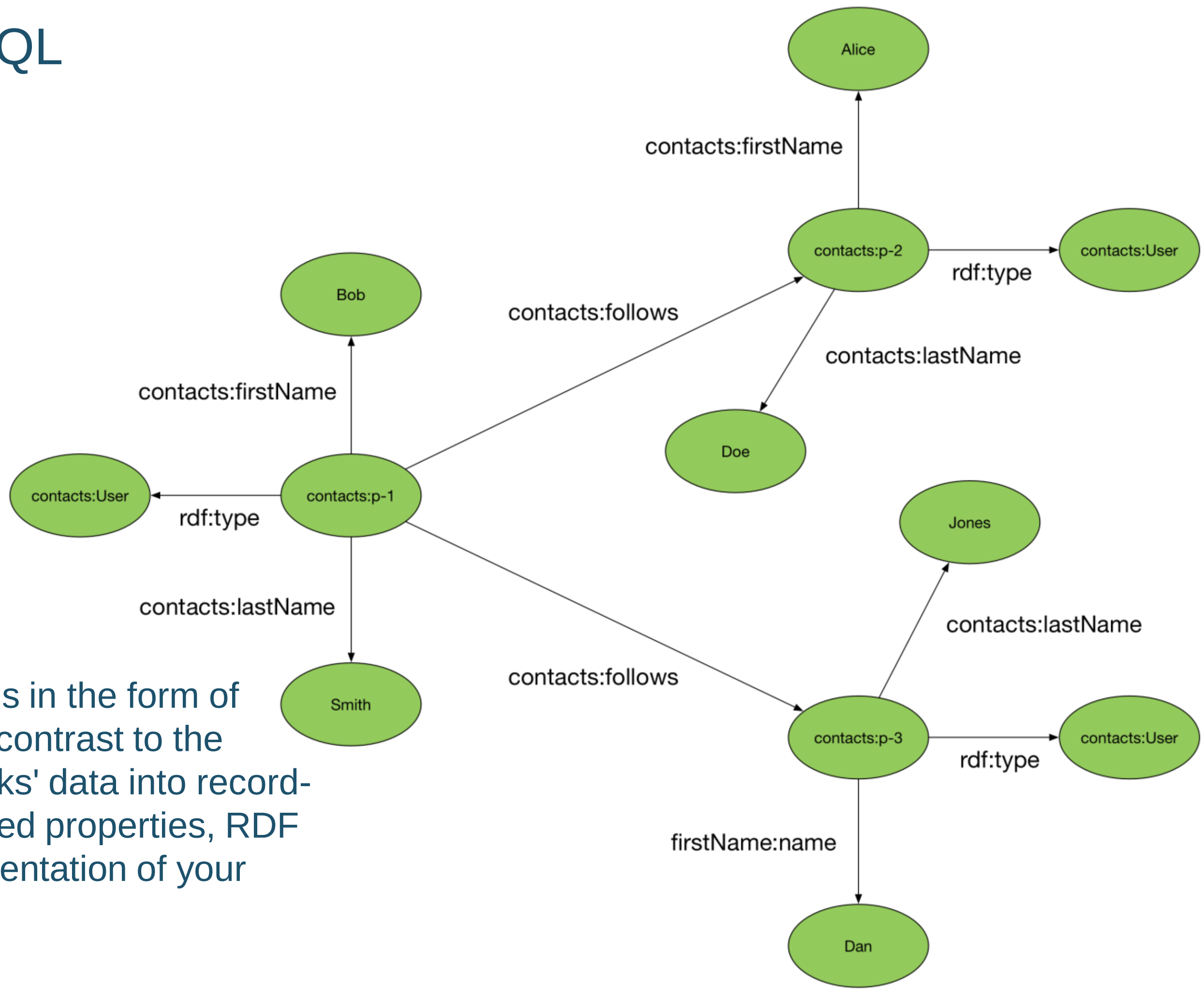
You can attach one or more properties to each of the vertices and edges in your graph. Typically, you use vertex properties to represent the attributes of entities in your domain, and edge properties to represent the strength, weight or quality of a relationship. You can also use properties to represent metadata – timestamps, access control lists, etc.

Different data models and query languages

@prefix contacts: <http://www.example.com/social#>

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

RDF Graph and SPARQL



RDF encodes resource descriptions in the form of subject-predicate-object triples. In contrast to the property graph model, which 'chunks' data into record-like vertices and edges with attached properties, RDF creates a more fine-grained representation of your domain.

Purpose-built databases: common categories and use cases

	Relational	Key-value	Document	In-memory	Graph	Time-series	Ledger
	Referential integrity, ACID transactions, schema-on-write	High throughput, low-latency reads & writes, endless scale	Store documents and quickly access querying on any attribute	Query by key with microsecond latency	Quickly and easily create and navigate relationships between data	Collect, store, and process data sequenced by time	Complete, immutable, and verifiable history of all changes to application data
Common Use Cases	Lift & shift, ERP, CRM, finance	Real-time bidding, shopping cart, social, product catalog, customer preferences	Content management, personalization, mobile	Leaderboards, real-time analytics, caching	Fraud detection, social networking, recommendation engine	IoT applications, event tracking	Systems of record, supply chain, health care, registrations, financial
AWS Service(s)	Aurora, RDS	DynamoDB	DocumentDB	ElastiCache	Neptune	Timestream	QLDB

Use-the-Right-Tool mantra persists across categories

AWS Compute services

Category	Use cases	AWS service
Virtual machines	Secure and resizable compute capacity (virtual servers) in the cloud	Amazon Elastic Compute Cloud (EC2)
	Easy-to-use cloud platform that offers you everything you need to build an application or website	Amazon Lightsail
Containers	Highly secure, reliable, and scalable way to run containers	Amazon Elastic Container Service (ECS)
	Easily store, manage, and deploy container images	Amazon Elastic Container Registry (ECR)
	Fully managed Kubernetes service	Amazon Elastic Kubernetes Service (EKS)
Serverless	Run code without thinking about servers. Pay only for the compute time you consume	AWS Lambda
	Serverless compute for containers	AWS Fargate
Edge and hybrid	Run AWS infrastructure and services on premises for a truly consistent hybrid experience	AWS Outposts
	Delivery ultra-low latency application for 5G devices	AWS Wavelength
	Rapidly extend, migrate, and protect your VMware environment to the AWS cloud	VMware Cloud on AWS
	Run latency sensitive applications closer to end-users	AWS Local Zones
Cost and capacity management	Run fault-tolerant workload for up to 90% off	Amazon EC2 Spot Instances
	Automatically add or remove compute capacity to meet changes in demand	Amazon EC2 Autoscaling
	Fully managed batch processing at any scale	AWS Batch

AWS Security, Identity, & Compliance services

Category	Use cases	AWS service
Identity & access management	Identity management for your apps	Amazon Cognito
	Managed Microsoft Active Directory	AWS Directory Service
	Manage user access and encryption keys	AWS Identity & Access Management (IAM)
	Simple, secure service to share AWS resources	AWS Resource Access Manager
	Cloud single-sign-on (SSO) service	AWS Single Sign-On
Detective controls	Unified security and compliance center	AWS Security Hub
	Managed threat detection service	Amazon GuardDuty
	Analyze application security	Amazon Inspector
	Investigate potential security issues	Amazon Detective
Infrastructure protection	DDoS protection	AWS Shield
	Filter malicious web traffic	AWS Web Application Firewall (WAF)
	Central management of firewall rules	AWS Firewall Manager
Data protection	Discover, classify and protect your data	Amazon Macie
	Key storage and management	AWS Key Management Service (KMS)
	Hardware based key storage for regulatory compliance	AWS CloudHSM
	Provision, manage, and deploy public and private SSL/TLS	AWS Certificate Manager

AWS Cloud Storage Products

If You Need:	Consider Using:
Persistent local storage for Amazon EC2, for relational and NoSQL databases, data warehousing, enterprise applications, Big Data processing, or backup and recovery	Amazon Elastic Block Store (Amazon EBS)
A simple, scalable, elastic file system for Linux-based workloads for use with AWS Cloud services and on-premises resources. It is built to scale on demand to petabytes without disrupting applications, growing and shrinking automatically as you add and remove files, so your applications have the storage they need – when they need it.	Amazon Elastic File System (Amazon EFS)
A fully managed file system that is optimized for compute-intensive workloads, such as high performance computing, machine learning, and media data processing workflows, and is seamlessly integrated with Amazon S3	Amazon FSx for Lustre
A fully managed native Microsoft Windows file system built on Windows Server so you can easily move your Windows-based applications that require file storage to AWS, including full support for the SMB protocol and Windows NTFS, Active Directory (AD) integration, and Distributed File System (DFS).	Amazon FSx for Windows File Server
A scalable, durable platform to make data accessible from any Internet location, for user-generated content, active archive, serverless computing, Big Data storage or backup and recovery	Amazon Simple Storage Service (Amazon S3)
Highly affordable long-term storage classes that can replace tape for archive and regulatory compliance	Amazon S3 Glacier & Amazon S3 Glacier Deep Archive
A hybrid storage cloud augmenting your on-premises environment with Amazon cloud storage, for bursting, tiering or migration	AWS Storage Gateway
A portfolio of services to help simplify and accelerate moving data of all types and sizes into and out of the AWS cloud	Cloud Data Migration Services
A fully managed backup service that makes it easy to centralize and automate the back up of data across AWS services in the cloud as well as on premises using the AWS Storage Gateway.	AWS Backup

ML Services
Build, train, and deploy ML fast

AI Services
Easily add intelligence to your applications

Frameworks
Choice and flexibility with broadest framework support

Compute
Fastest and lowest-cost compute options

Analytics and Security
Comprehensive capabilities, no compromise

Learning Tools
Get deep on ML with AWS DeepRacer and DeepLens

AI SERVICES

VISION	SPEECH		TEXT			SEARCH	CHATBOTS	PERSONALIZATION	FORECASTING	FRAUD	DEVELOPMENT	CONTACT CENTERS
Amazon Rekognition	Amazon Polly	Amazon Transcribe *+Medical*	Amazon Comprehend *+Medical*	Amazon Translate	Amazon Textract	Amazon Kendra	Amazon Lex	Amazon Personalize	Amazon Forecast	Amazon Fraud Detector	Amazon CodeGuru	Contact Lens *For Amazon Connect*

ML SERVICES

Amazon SageMaker	Ground Truth	AWS Marketplace for ML	SageMaker Studio IDE									Neo	Augmented AI
			Built-in algorithms	Notebooks	Experiments	Processing	Model training & tuning	Debugger	Autopilot	Model hosting	Model Monitor		

ML FRAMEWORKS & INFRASTRUCTURE

TensorFlow, mxnet, PyTorch, Gluon, Keras, scikit-learn, Horovod, DeepGraphLibrary

Deep Learning AMIs & Containers	GPUs & CPUs	Elastic Inference	Inferentia	FPGA

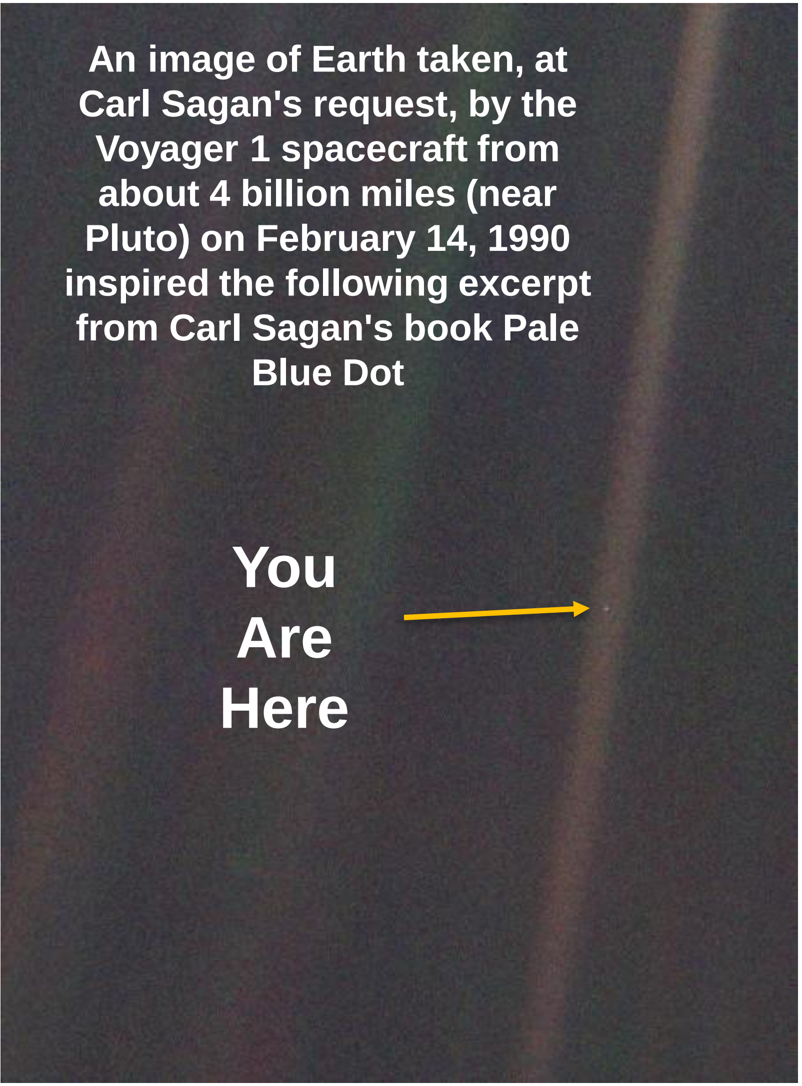
Jerry Hargrove | Periodic Table of Amazon Web Services

The 2030 Agenda for Sustainable Development, adopted by all United Nations Member States in 2015, provides a shared blueprint for peace and prosperity for people and the planet, now and into the future.

At its heart are the 17 Sustainable Development Goals (SDGs), which are an urgent call for action by all countries - developed and developing - in a global partnership.

We have less than 10 years to solve the SDGs and AI holds great promise.

AI on AWS is helping with all of these



Look again at that dot. That's here. That's home. That's us.

On it everyone you love, everyone you know, everyone you ever heard of, every human being who ever was, lived out their lives.

The aggregate of our joy and suffering, thousands of confident religions, ideologies, and economic doctrines, every hunter and forager, every hero and coward, every creator and destroyer of civilization, every king and peasant, every young couple in love, every mother and father, hopeful child, inventor and explorer, every teacher of morals, every corrupt politician, every "superstar," every "supreme leader," every saint and sinner in the history of our species lived there –

on a mote of dust suspended in a sunbeam.

Carl Sagan, Pale Blue Dot, 1994

Questions

Stream Data Processing

https://harvard-iacs.github.io/2020-CS205/