FRANZ INC.
Web 3.0's Database

AllegroGraph in the Elastic Cloud (Amazon, EC2)

Jans Aasman
Franz Inc
(info@franz.com)

This talk

- AllegroGraph in 5 minutes
- EC2
- Running AllegroGraph on EC2
- An example with 10 Billion triples
  - The dataset
  - Loading
  - Federation and Queries
- Future work

What is AllegroGraph [1]

- Scalable and persistent quadstore
- Federated
  - Create an abstract store that is a collection of other triple stores. Prolog and SPARQL and Reasoning work transparently against abstract store
- Compliant with standards
  - RDF, RDFS, OWL, SPARQL, Named Graphs, ISO Prolog, OWL-lite reasoning

What is AllegroGraph [2]

- Relational database efficiency for range queries
  - We support most xml schema types (dates, times, longitudes, latitudes, durations, telephone numbers, etc)
- Spatial database efficiency for geospatial primitives
  - Find elements in bounding boxes as fast as in spatial databases
- Temporal reasoning
  - Reasoning about times and intervals (Allen Logic)
- Social Network Analytics library
  - Find actor degrees and centrality, cliques, group centrality and cohesiveness

Activity Recognition

- **Our customers use AllegroGraph as an event database with social network analysis and geospatial and temporal reasoning**

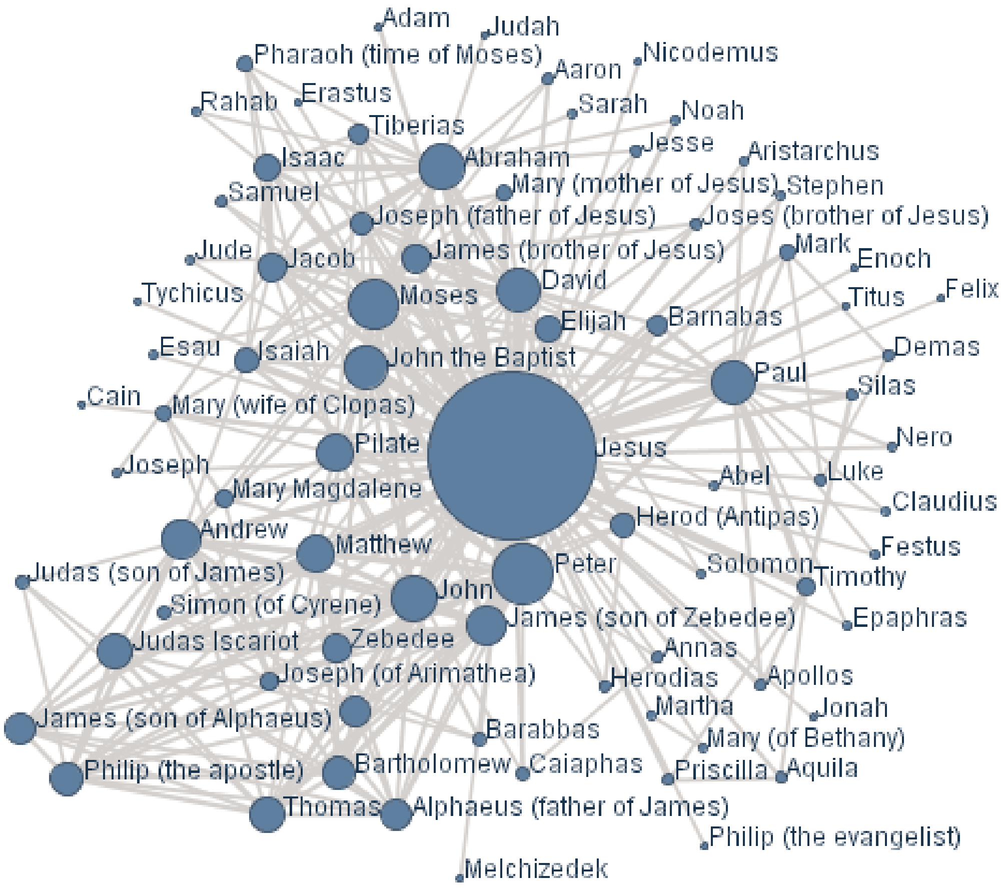
Find all meetings that happened in December within 5 miles of Berkeley that was attended by the most important person in Jans' friends and friends of friends.

```
(select (?x)
   (ego-group !person:jans knows ?group 2)            SNA
   (actor-centrality-members ?group knows ?x ?num)    SNA
   (q ?event !fr:actor ?x)                            DB Lookup
   (qs ?event !rdf:type !fr:Meeting)                  RDFS
   (interval-during ?event 2007-12-01 2007-12-31)     Temporal
   (geo-box-around !geoname:Berkeley ?event 5 miles)  Spatial
   !)
```

FRANZ INC.
Web 3.0's Database

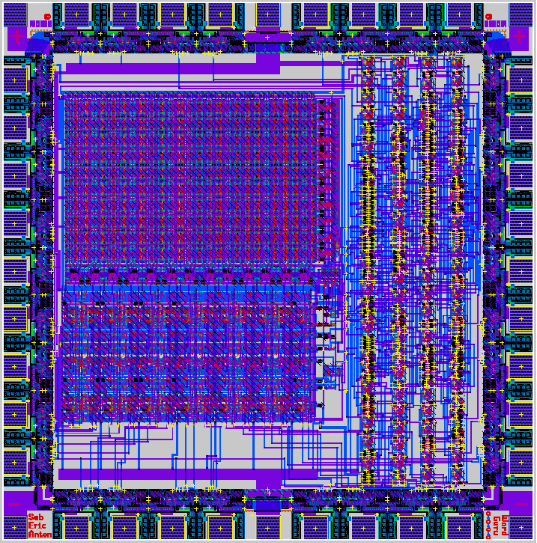
Social Network Analysis Answers 4 questions

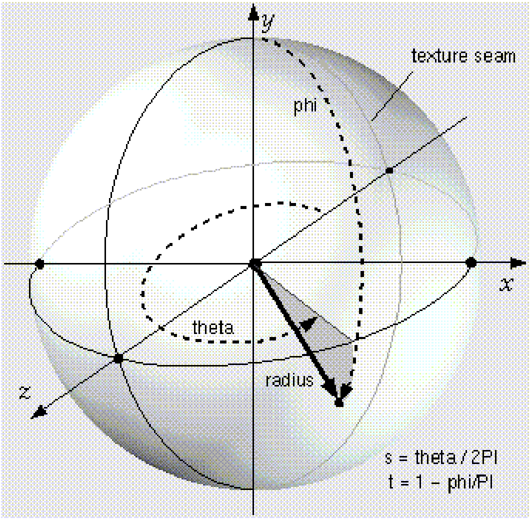
- How far is P1 from P2 (and how strong is the relation?)
- To what groups does this person belong (ego groups, cliques?)
- How important is this person in the group?
- Does this group have a leader, how cohesive are they?



GeoSpatial

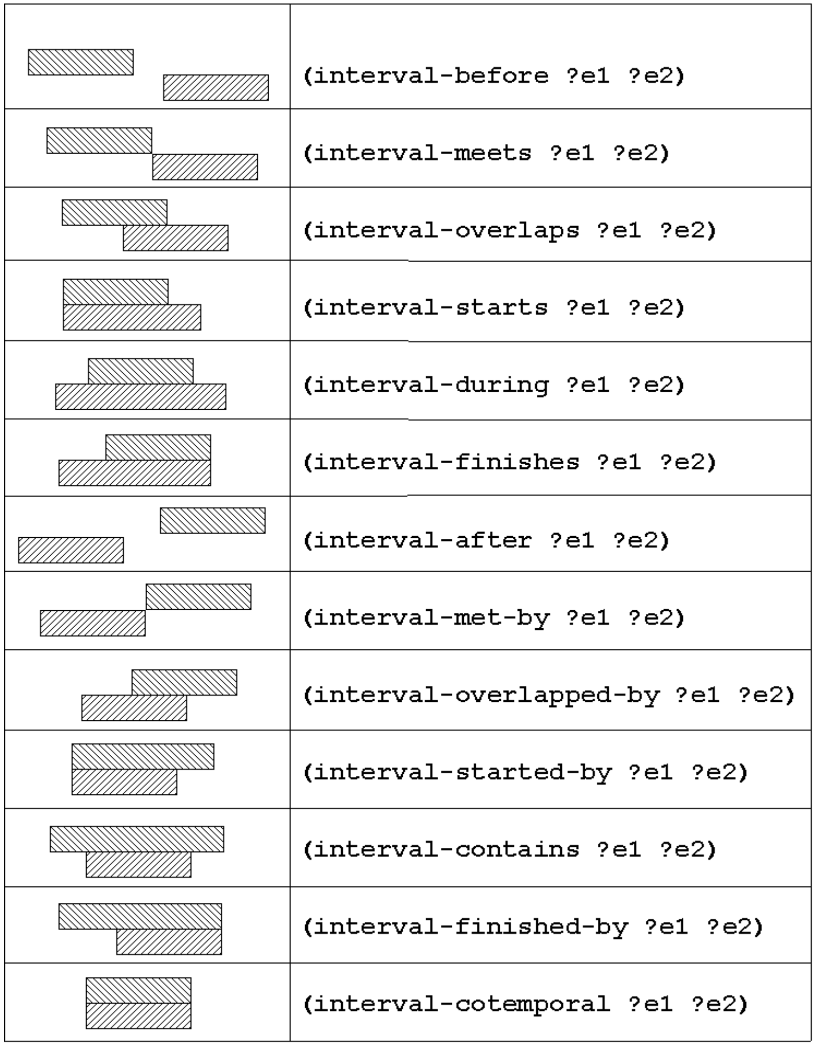
- Make the following super efficient
  - Where did something happen?
  - How far was event1 from event2?
  - Find all the events that occurred in a bounding box or radius of M miles?
  - Do these two shapes overlap?
  - Find all the objects in the intersection of two shapes
- On a very large scale
  - when things don't fit in memory
  - millions of events and polygons





Temporal Reasoning

- Adhere to our convention to encode StartTimes and EndTimes and enjoy efficient temporal primitives
- Implementation of Allen's interval logic primitives



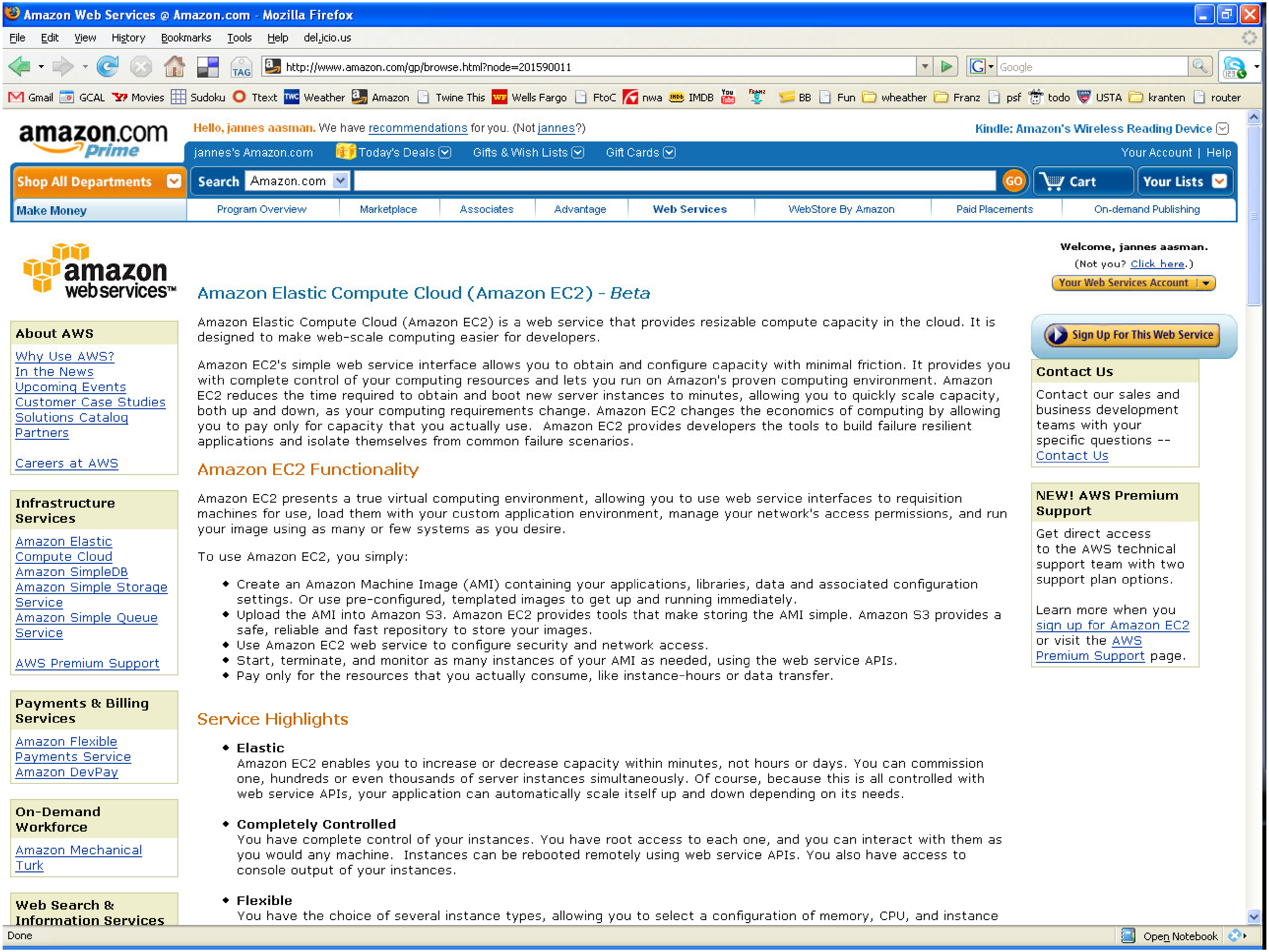
```
(interval-before ?e1 ?e2)
(interval-meets ?e1 ?e2)
(interval-overlaps ?e1 ?e2)
(interval-starts ?e1 ?e2)
(interval-during ?e1 ?e2)
(interval-finishes ?e1 ?e2)
(interval-after ?e1 ?e2)
(interval-met-by ?e1 ?e2)
(interval-overlapped-by ?e1 ?e2)
(interval-started-by ?e1 ?e2)
(interval-contains ?e1 ?e2)
(interval-finished-by ?e1 ?e2)
(interval-cotemporal ?e1 ?e2)
```

Web 3.0's Database

But what if I have a 10 Billion Triples

But what if I have a 10 Billion Triples

- Netezza
  - $ 500,000 to 1,500,000
- Large memory machine
  - $ 200,000 to 1,000,000
- Build your own cluster
  - $ 50,000 to 100,000
- Amazon’s EC2 to the rescue
  - $ 192 for two days


Amazon Web Services @ Amazon.com - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
http://www.amazon.com/gp/browse.html?node=201590011
Google
Gmail GCAL Movies Sudoku Ttext Weather Amazon Twine This Wells Fargo FtoC nwa IMDB BB Fun wheather Franz psf todo USTA kranten router
amazon.com Prime
Hello, jannes aasman. We have recommendations for you. (Not jannes?)
Kindle: Amazon's Wireless Reading Device
jannes's Amazon.com
Today's Deals
Gifts & Wish Lists
Gift Cards
Your Account | Help
Shop All Departments
Search
Amazon.com
GO
Cart
Your Lists
Make Money
Program Overview
Marketplace
Associates
Advantage
Web Services
WebStore By Amazon
Paid Placements
On-demand Publishing
amazon webservices™
Welcome, jannes aasman.
(Not you? Click here.)
Your Web Services Account
Sign Up For This Web Service
Done
Open Notebook

About AWS

Why Use AWS?
In the News
Upcoming Events
Customer Case Studies
Solutions Catalog
Partners

Careers at AWS

Infrastructure Services

Amazon Elastic Compute Cloud
Amazon SimpleDB
Amazon Simple Storage Service
Amazon Simple Queue Service

AWS Premium Support

Payments & Billing Services

Amazon Flexible Payments Service
Amazon DevPay

On-Demand Workforce

Amazon Mechanical Turk

Web Search & Information Services

Amazon Elastic Compute Cloud (Amazon EC2) - *Beta*

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.

Amazon EC2's simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use. Amazon EC2 provides developers the tools to build failure resilient applications and isolate themselves from common failure scenarios.

Amazon EC2 Functionality

Amazon EC2 presents a true virtual computing environment, allowing you to use web service interfaces to requisition machines for use, load them with your custom application environment, manage your network's access permissions, and run your image using as many or few systems as you desire.

To use Amazon EC2, you simply:

- Create an Amazon Machine Image (AMI) containing your applications, libraries, data and associated configuration settings. Or use pre-configured, templated images to get up and running immediately.
- Upload the AMI into Amazon S3. Amazon EC2 provides tools that make storing the AMI simple. Amazon S3 provides a safe, reliable and fast repository to store your images.
- Use Amazon EC2 web service to configure security and network access.
- Start, terminate, and monitor as many instances of your AMI as needed, using the web service APIs.
- Pay only for the resources that you actually consume, like instance-hours or data transfer.

Service Highlights

- **Elastic**
  Amazon EC2 enables you to increase or decrease capacity within minutes, not hours or days. You can commission one, hundreds or even thousands of server instances simultaneously. Of course, because this is all controlled with web service APIs, your application can automatically scale itself up and down depending on its needs.

- **Completely Controlled**
  You have complete control of your instances. You have root access to each one, and you can interact with them as you would any machine. Instances can be rebooted remotely using web service APIs. You also have access to console output of your instances.

- **Flexible**
  You have the choice of several instance types, allowing you to select a configuration of memory, CPU, and instance

Contact Us

Contact our sales and business development teams with your specific questions -- Contact Us

NEW! AWS Premium Support

Get direct access to the AWS technical support team with two support plan options.

Learn more when you sign up for Amazon EC2 or visit the AWS Premium Support page.

EC2 - advantages

- Scalability in two dimensions
  - Use as many machines as you need
  - Various machine sizes available
- High availability
- High bandwidth
- No upfront investments
- Our example with 10 Billion triples cost us (10 * 2 * 24 * 0.40 =) 192 dollars.

EC2 – current problems

- When turning off instances you lose the local data
- Copying huge amounts of data to S3 is clumsy and a pain
- HOWEVER: will be fixed this year

Running AllegroGraph on EC2

- Running AllegroGraph on EC2
  - http://agraph.franz.com/ec2.lhtml

Loading a 10 billion triples

- 1 billion telecom CDRs in csv files, 10 fields per line
- Take 6 hours to upload to ec2

- 10 ‘large’ instances
- 100 M CDRs per instance (= 1 B triples)
- 4 parallel loads on each instance
  - loading 250 M triples per load

```
Total load time:        2:59 hours
Total indexing time:    3:20 hours
Total time :            6.19 hours
```

Find all calls from a to b

• On one 250 M TS	• 0.002 secs
• On a federation of 4 TS on one machine	• 0.008 secs
• On the 10 B set on 10 machines	• 0.083 secs

The future

- Joins suffer from too many triple stores
- Experiment with informed federation
  - Over time
  - Per predicate
  - Per type of object
- Expect solutions and patterns from Franz in 2008
- Who wants this as a service?
  - info@franz.com