Modern Graph Analysis using Tinkerpop and Janusgraph

Pedro Pires
Ricardo Faria

March 24, 2018

Know the unknown.

WeDo technologies

AGENDA

- Introduction
- Goal
- Graphs
- Tools
- Hands on
- Conclusion

Introduction

Hi!

RAID.Cloud Development Lead at
WeDo Technologies, Braga

linkedin.com/in/rgomesf/

RAID.Cloud ML Lead at
WeDo Technologies, Braga

linkedin.com/in/pmpires/

WeDo Technologies - Quick Profile

CUSTOMERS, TEAM AND CULTURE

More than
220 CUSTOMERS
in more than 100 countries

Offices in
10 COUNTRIES
and in 5 continents

A team of 600+ people
from more than
20 NATIONALITIES

A **“WEDO” CULTURE**

Proud of being
part of this
COMMUNITY!

STRATEGY AND MARKET PRESENCE

# 1 IN THE WORLD
in Telecom Revenue Assurance and Fraud Management Software

Gartner

Stratecast / Frost & Sullivan

Analysis Mason

World class reference customers
in Telecom, Retail, Energy, Healthcare and Financial Industries

INDIRECT CHANNEL STRATEGY
has successfully started with two global/Worldwide partners already certified

WeDo Technologies - Offer

SOFTWARE HOUSE COMPLEMENTED WITH BUSINESS CONSULTING EXPERTISE

SOFTWARE

Products covering Revenue Assurance and Fraud Management and niche Business Challenges in the Telecom industry.

MANAGED SERVICES

Our Managed Services addresses key issues that impact Risk Management activities namely cost-reduction, skills acquisition, and processes improvement.

PROFESSIONAL SERVICES

WeDo Technologies provides Professional Services for our Software and Solutions.

WeDo Technologies - Product Portfolio

COMMON ARCHITECTURE FOR ALL SOFTWARE PRODUCTS

RAID RISK MANAGEMENT

RAID REVENUE ASSURANCE

RAID FRAUD MANAGEMENT

RAID BUSINESS ASSURANCE

ON PREMISES MODULES

CLOUD APPS

RAID BUSINESS MANAGEMENT

RAID INCENTIVES

RAID COLLECTIONS

RAID ROAMING

ON PREMISES MODULES

RAID BUSINESS OPTIMIZATION

RAID OPTIMIZE

ON PREMISES MODULES

WeDo Technologies - Product Portfolio

COMMON ARCHITECTURE FOR ALL SOFTWARE PRODUCTS

RAID RISK MANAGEMENT

RAID REVENUE ASSURANCE

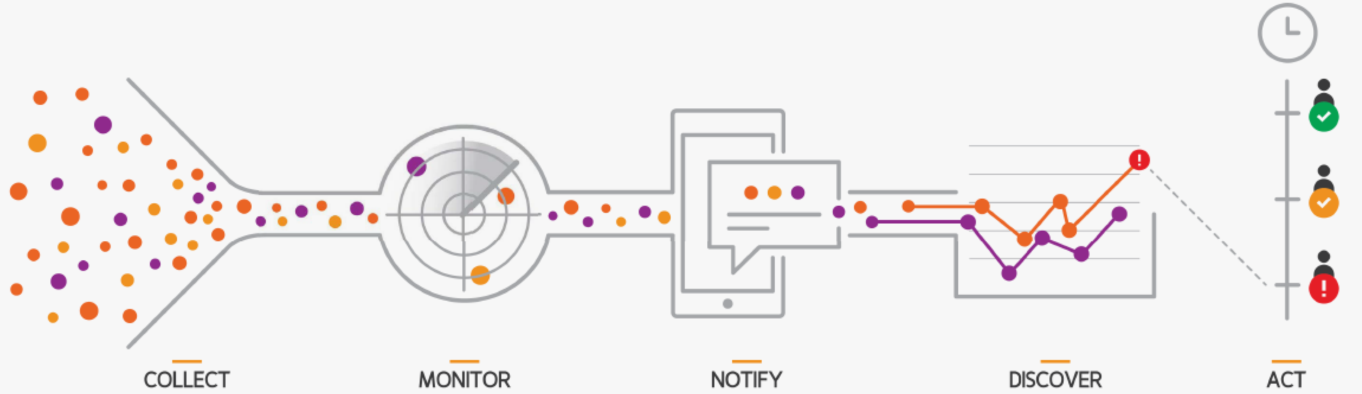
RAID FRAUD MANAGEMENT

RAID BUSINESS ASSURANCE

ON PREMISES MODULES

CLOUD APPS

NEXT GENERATION END-TO-END FRAUD MANAGEMENT SOFTWARE + OPTIONAL PREBUILT MODULES READY TO USE



- Roaming Fraud
- Bypass Fraud
- IRSF Fraud
- Assets Misappropriation
- PBX Hacking
- Prepaid Fraud
- High Usage Fraud
- Subscription Fraud
- Wholesale Fraud
- Dealers Fraud

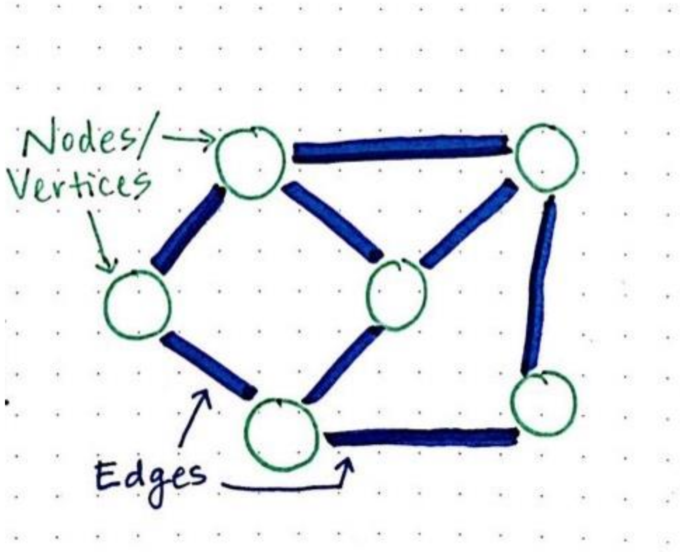
http://www.wedotechnologies.com/en/careers/

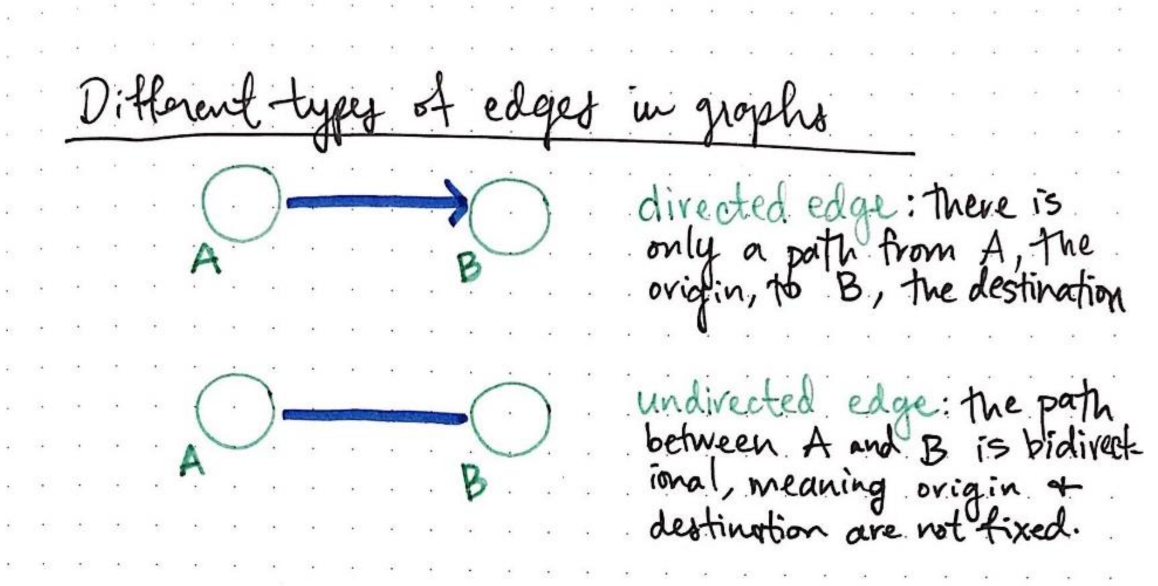
Goal

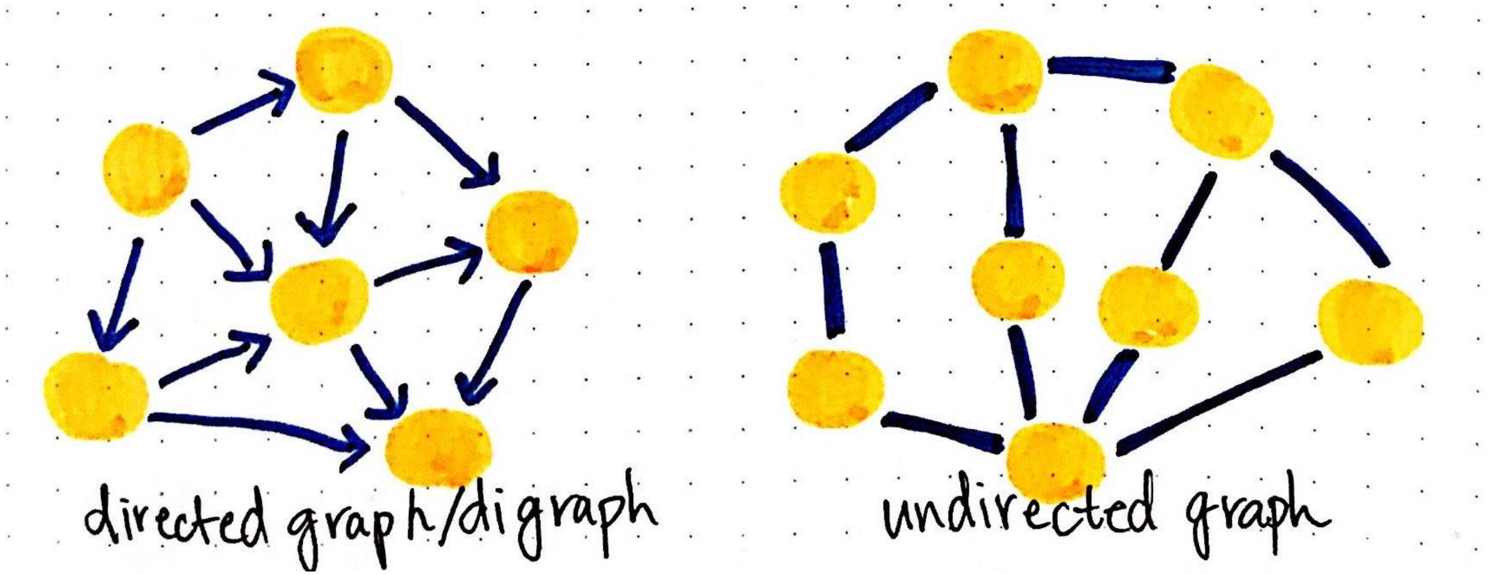
Show the potential of Graph DB

Graphs

Concepts



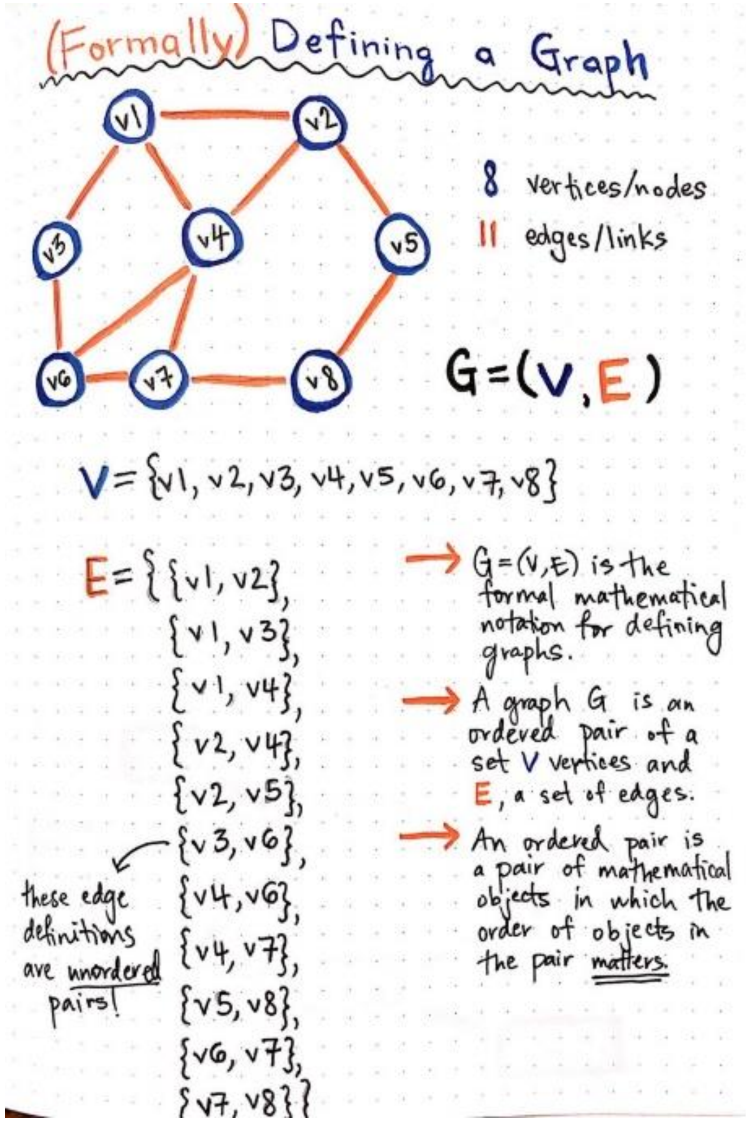

Different types of edges in graphs
A
B
directed edge: there is only a path from A, the origin, to B, the destination
A
B
undirected edge: the path between A and B is bidirectional, meaning origin + destination are not fixed.



https://medium.com/basecs/a-gentle-introduction-to-graph-theory-77969829ead8

Graphs

Concepts



(Formally) Defining a Graph

8 vertices/nodes

11 edges/links

$G=(V,E)$

$V=\{v1, v2, v3, v4, v5, v6, v7, v8\}$

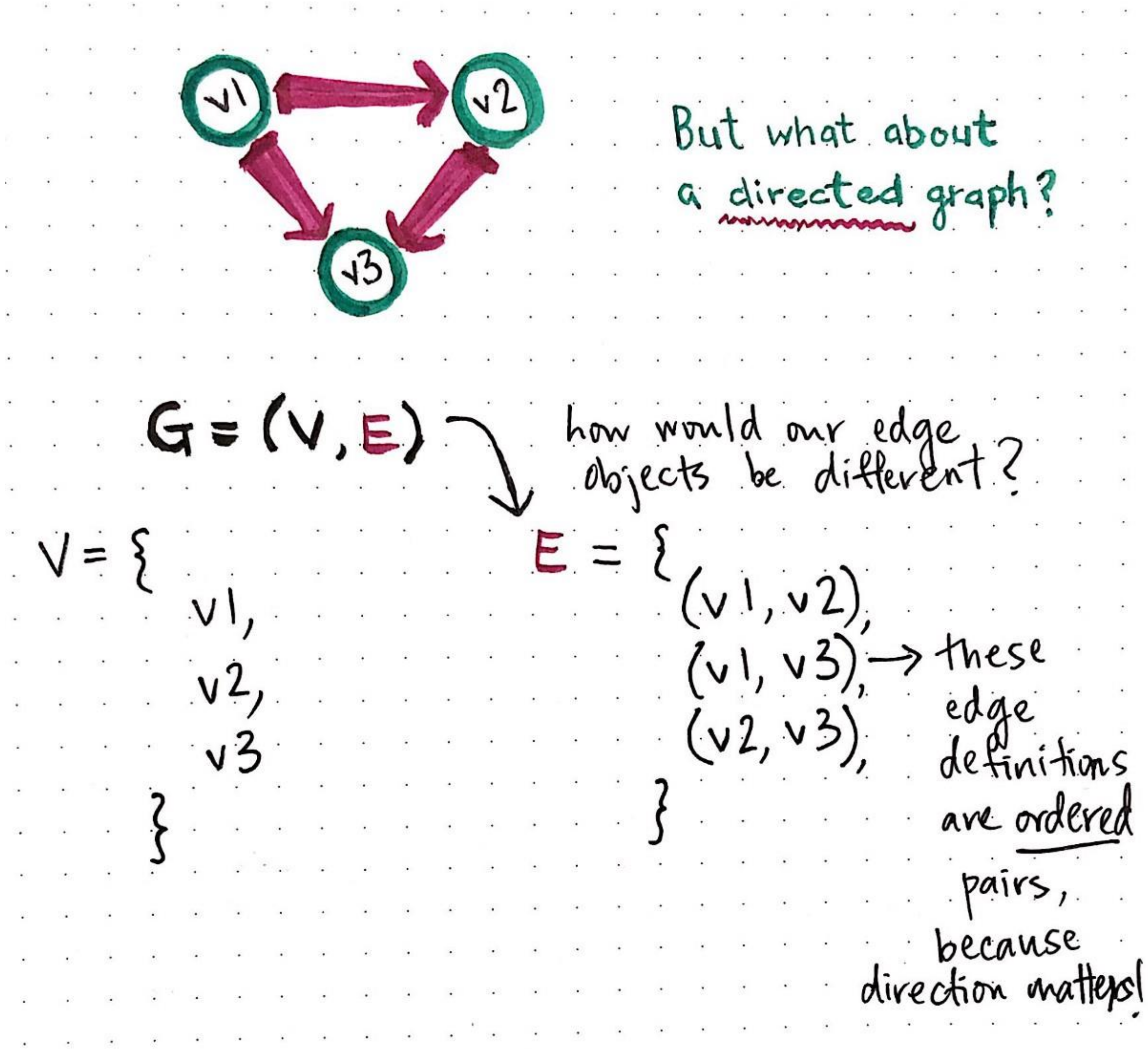
$E=\{\{v1, v2\}, \{v1, v3\}, \{v1, v4\}, \{v2, v4\}, \{v2, v5\}, \{v3, v6\}, \{v4, v6\}, \{v4, v7\}, \{v5, v8\}, \{v6, v7\}, \{v7, v8\}\}$

these edge definitions are unordered pairs!

→ $G=(V,E)$ is the formal mathematical notation for defining graphs.

→ A graph G is an ordered pair of a set V vertices and E, a set of edges.

→ An ordered pair is a pair of mathematical objects in which the order of objects in the pair matters.



But what about a directed graph?

$G=(V,E)$ how would our edge objects be different?

$V=\{v1, v2, v3\}$

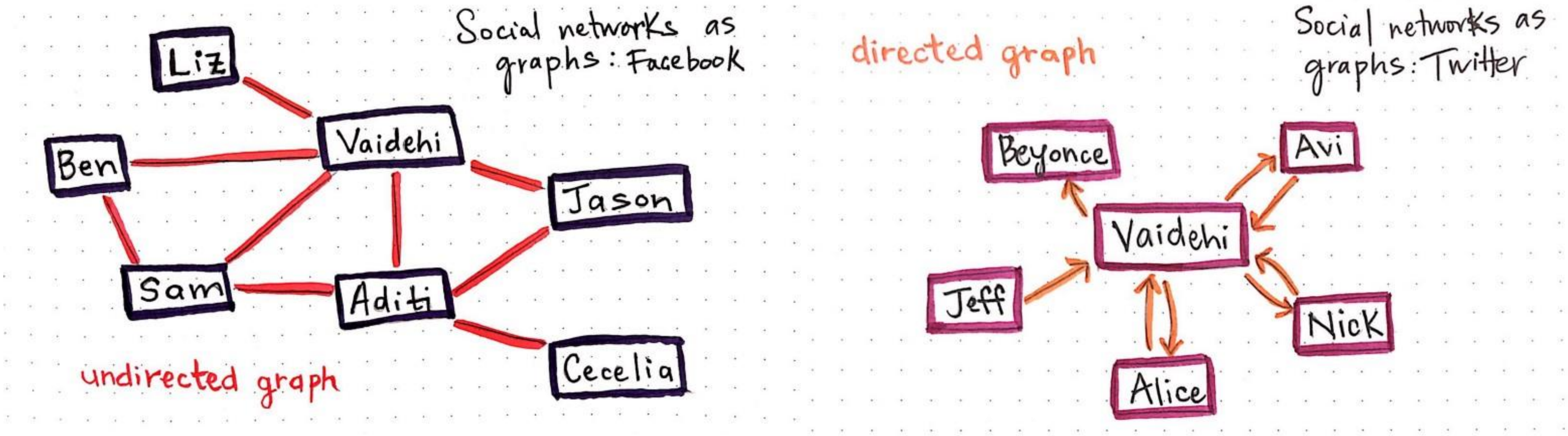
$E=\{(v1, v2), (v1, v3), (v2, v3),\}$

→ these edge definitions are ordered pairs, because direction matters!

https://medium.com/basecs/a-gentle-introduction-to-graph-theory-77969829ead8

Graphs

Concepts



https://medium.com/basecs/a-gentle-introduction-to-graph-theory-77969829ead8

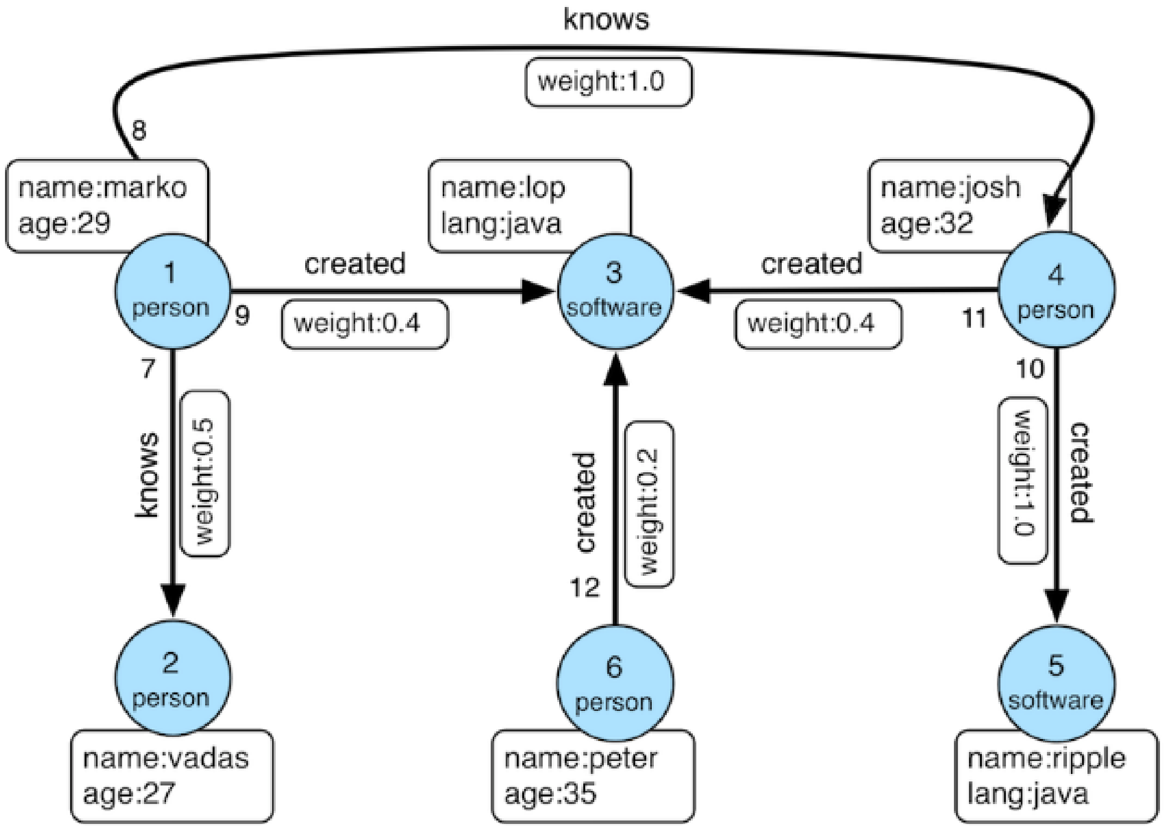
Graphs

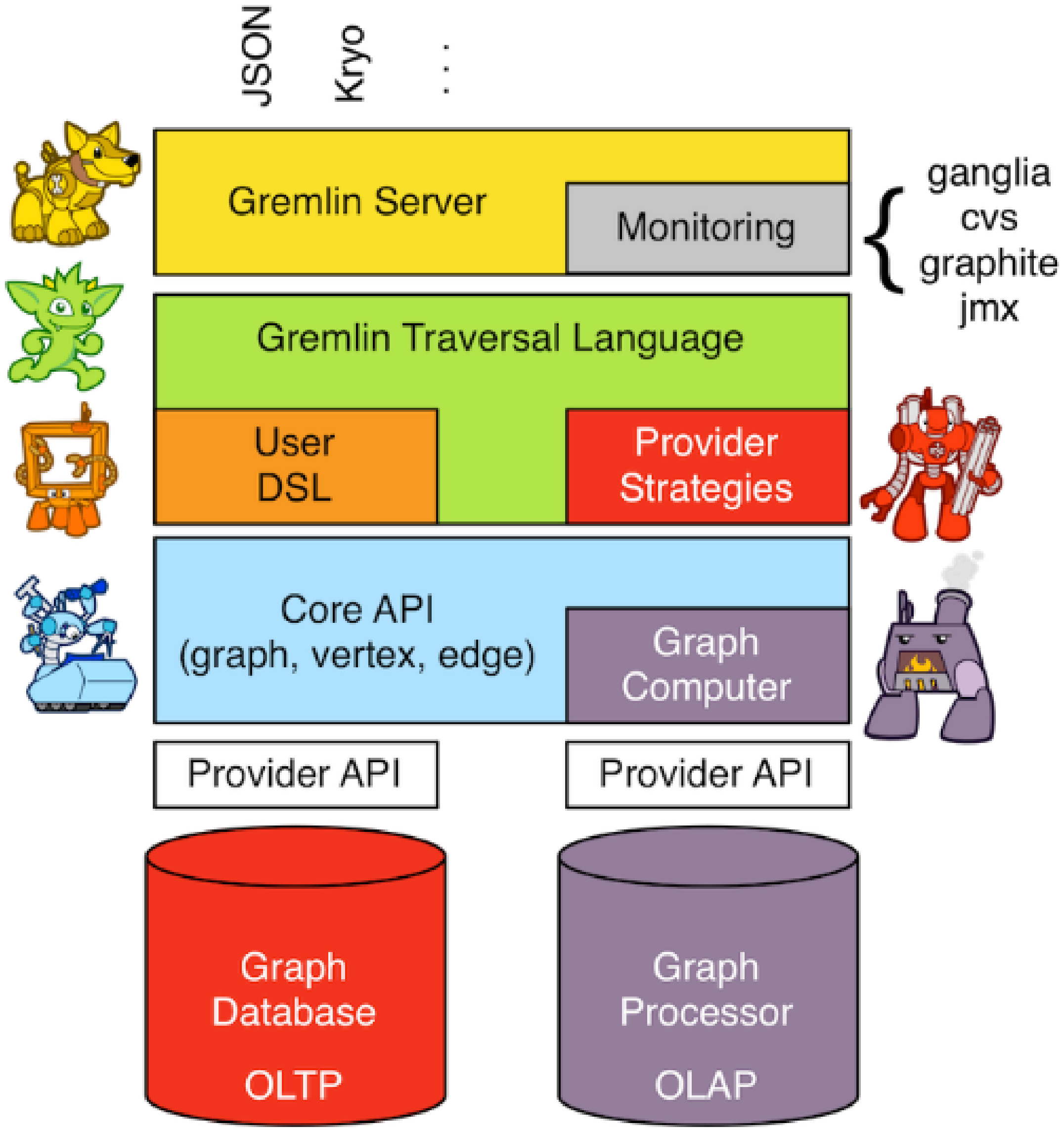
Concepts

https://www.youtube.com/watch?v=TwHy2DuWB3k

Graphs

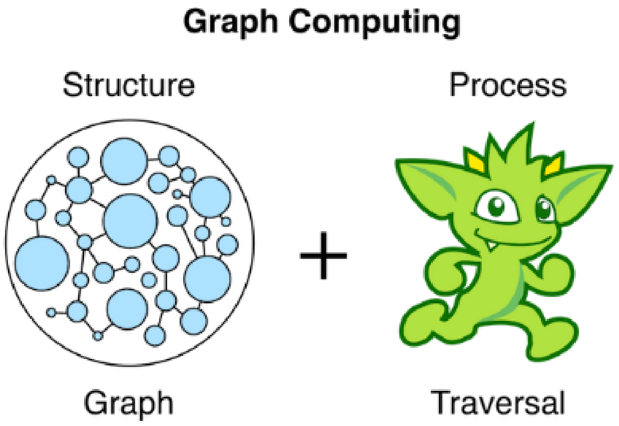
Apache TinkerPop



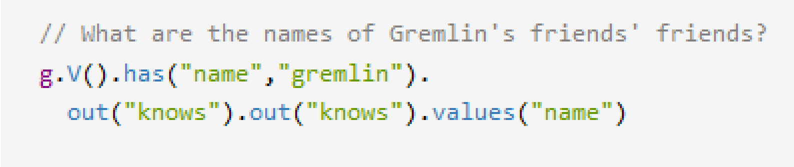


Graph Computing

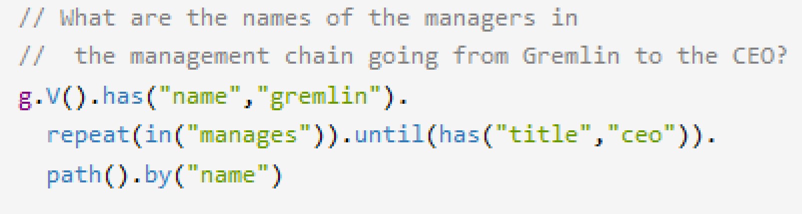
Gremlin Language



- Gremlin is useful for manually working with your graph
- Gremlin allows you to query a graph
- Gremlin can express complex graph traversals succinctly
- Gremlin is useful for exploring and learning about graphs
- Gremlin allows you to explore the Semantic Web/Web of Data
- Gremlin ensures that you are not tied to a particular graph backend
- Gremlin allows for universal path-based computations
- Gremlin is extensible and can be oriented to your particular use case
- Gremlin uses the Java API
- Gremlin is embedded in various JVM languages
- Gremlin is Turing complete



```
// What are the names of Gremlin's friends' friends?
g.V().has("name","gremlin").
  out("knows").out("knows").values("name")
```

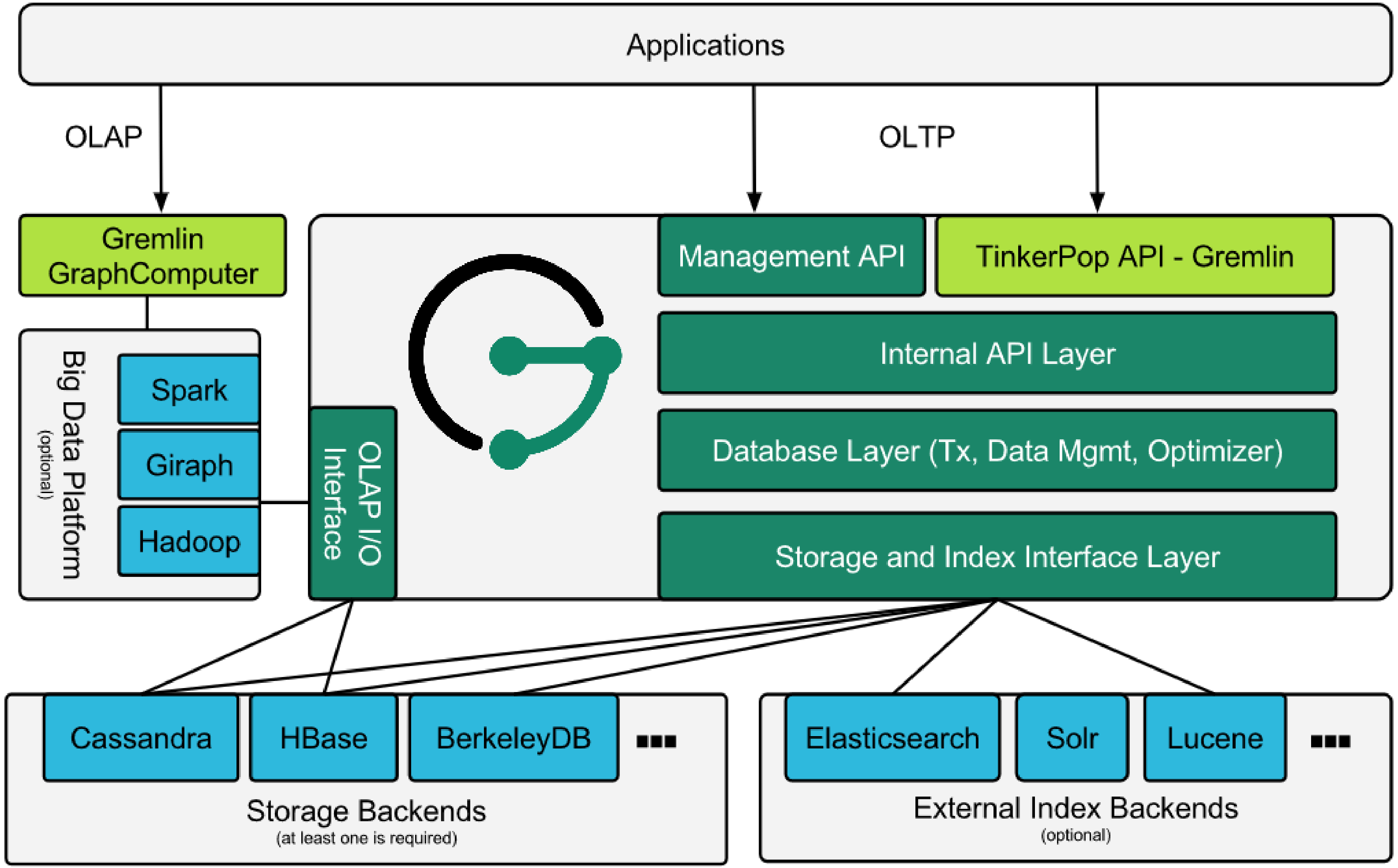


```
// What are the names of the managers in
//  the management chain going from Gremlin to the CEO?
g.V().has("name","gremlin").
  repeat(in("manages")).until(has("title","ceo")).
  path().by("name")
```

https://github.com/tinkerpop/gremlin/wiki/The-Benefits-of-Gremlin

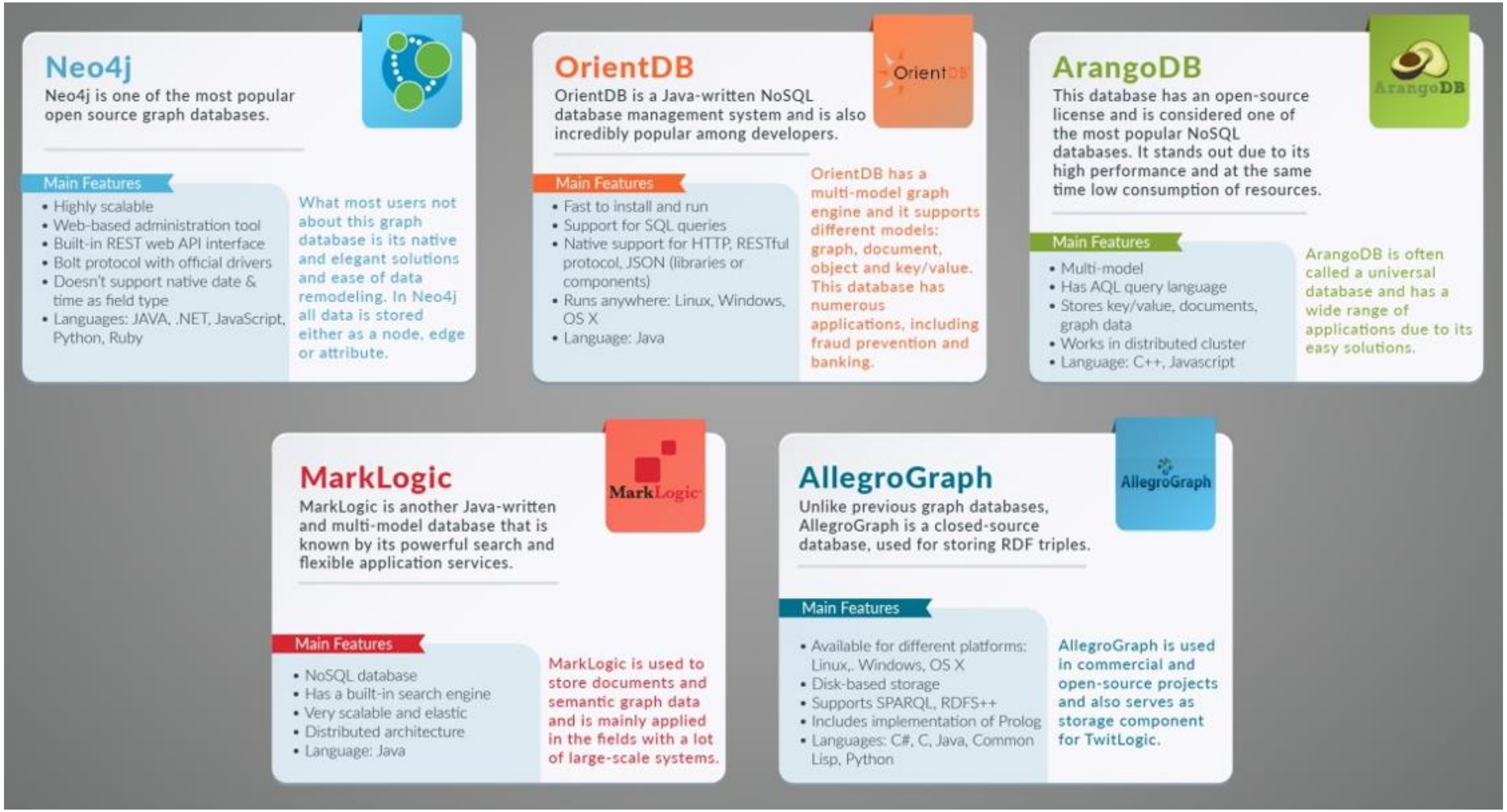
Tools

JanusGraph



Tools

Alternatives



Neo4j

Neo4j is one of the most popular open source graph databases.

Main Features

- Highly scalable
- Web-based administration tool
- Built-in REST web API interface
- Bolt protocol with official drivers
- Doesn't support native date & time as field type
- Languages: JAVA, .NET, JavaScript, Python, Ruby

What most users not about this graph database is its native and elegant solutions and ease of data remodeling. In Neo4j all data is stored either as a node, edge or attribute.

OrientDB

OrientDB is a Java-written NoSQL database management system and is also incredibly popular among developers.

Main Features

- Fast to install and run
- Support for SQL queries
- Native support for HTTP, RESTful protocol, JSON (libraries or components)
- Runs anywhere: Linux, Windows, OS X
- Language: Java

OrientDB has a multi-model graph engine and it supports different models: graph, document, object and key/value. This database has numerous applications, including fraud prevention and banking.

ArangoDB

This database has an open-source license and is considered one of the most popular NoSQL databases. It stands out due to its high performance and at the same time low consumption of resources.

Main Features

- Multi-model
- Has AQL query language
- Stores key/value, documents, graph data
- Works in distributed cluster
- Language: C++, Javascript

ArangoDB is often called a universal database and has a wide range of applications due to its easy solutions.

MarkLogic

MarkLogic is another Java-written and multi-model database that is known by its powerful search and flexible application services.

Main Features

- NoSQL database
- Has a built-in search engine
- Very scalable and elastic
- Distributed architecture
- Language: Java

MarkLogic is used to store documents and semantic graph data and is mainly applied in the fields with a lot of large-scale systems.

AllegroGraph

Unlike previous graph databases, AllegroGraph is a closed-source database, used for storing RDF triples.

Main Features

- Available for different platforms: Linux,. Windows, OS X
- Disk-based storage
- Supports SPARQL, RDFS++
- Includes implementation of Prolog
- Languages: C#, C, Java, Common Lisp, Python

AllegroGraph is used in commercial and open-source projects and also serves as storage component for TwitLogic.

https://dashbouquet.com/blog/web-development/top-5-graph-databases

Hands On

Hands On

Docker

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

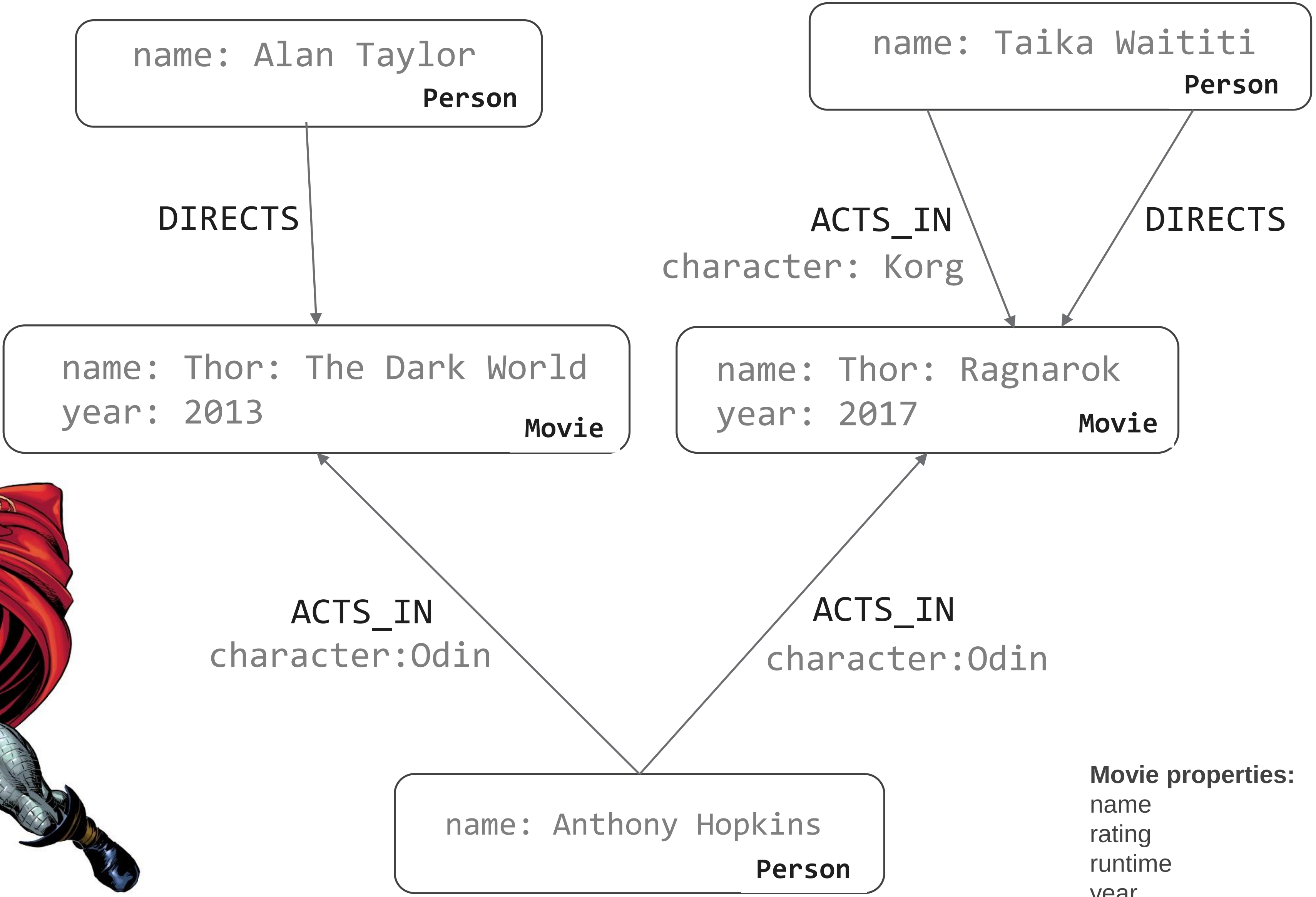
Hands On

1.1 - Docker

1. `git config --global core.autocrlf true`
2. `git clone https://github.com/rgomesf/janus.git`
3. `docker build --rm -f Dockerfile -t janus:latest .`
4. `docker run --rm --name=workshop -d -p 80:80 -p 8182:8182 janus`
5. `docker exec -it workshop bash`
6. `/work/janusgraph/bin/gremlin.sh`
7. `:remote connect tinkerpop.server conf/remote.yaml session-managed`
8. `:remote console`

Hands On

1.2 Schema



name: Alan Taylor
Person

DIRECTS

name: Taika Waititi
Person

ACTS_IN
character: Korg

DIRECTS

name: Thor: The Dark World
year: 2013
Movie

name: Thor: Ragnarok
year: 2017
Movie

ACTS_IN
character:Odin

ACTS_IN
character:Odin

name: Anthony Hopkins
Person

Movie properties:
name
rating
runtime
year

Hands On

1.3 Groovy – Load Graph

1. Check marvel.graphml contents.
2. Edit janus-inmemory-marvel.groovy to load the marvel.graphml
3. **https://github.com/rgomesf/janus**
4. **:q**
5. **exit**
6. docker stop workshop
7. docker build --rm -f Dockerfile -t janus:latest .
8. docker run --rm --name=workshop -d -p 80:80 -p 8182:8182 janus
9. docker exec -it workshop bash
10. /work/janusgraph/bin/gremlin.sh
11. :remote connect tinkerpop.server conf/remote.yaml session-managed
12. :remote console

http://docs.janusgraph.org/0.2.0/schema.html
http://docs.janusgraph.org/0.2.0/indexes.html

Hands On

1.4 Gremlin Language - Traversal

step	description
V	the vertex iterator of the graph (with key indices, V(key,value) possible)
E	the edge iterator of the graph (with key indices, E(key,value) possible)
out(labels...?)	out adjacent vertices to the vertex
outE(labels...?)	the outgoing edges of the vertex
in(labels...?)	in adjacent vertices to the vertex
inE(labels...?)	the incoming edges of the vertex
both(labels...?)	both adjacent vertices of the vertex
bothE(labels...?)	both incoming and outgoing edges of the vertex
outV	the outgoing tail vertex of the edge
inV	the incoming head vertex of the edge
bothV	both incoming and outgoing vertices of the edge
has(key)	emit the element if it has the property key
has(key,value)	allow element if has property
dedup()	emit only incoming objects that have not been seen before with optional closure being object to check on
groupBy(map?){closure}{closure}	emits input, but groups input after processing it by provided key-closure and value-closure
groupCount(map?){closure?}{closure?}	emits input, but updates a map for each input, where closures provides generic map update

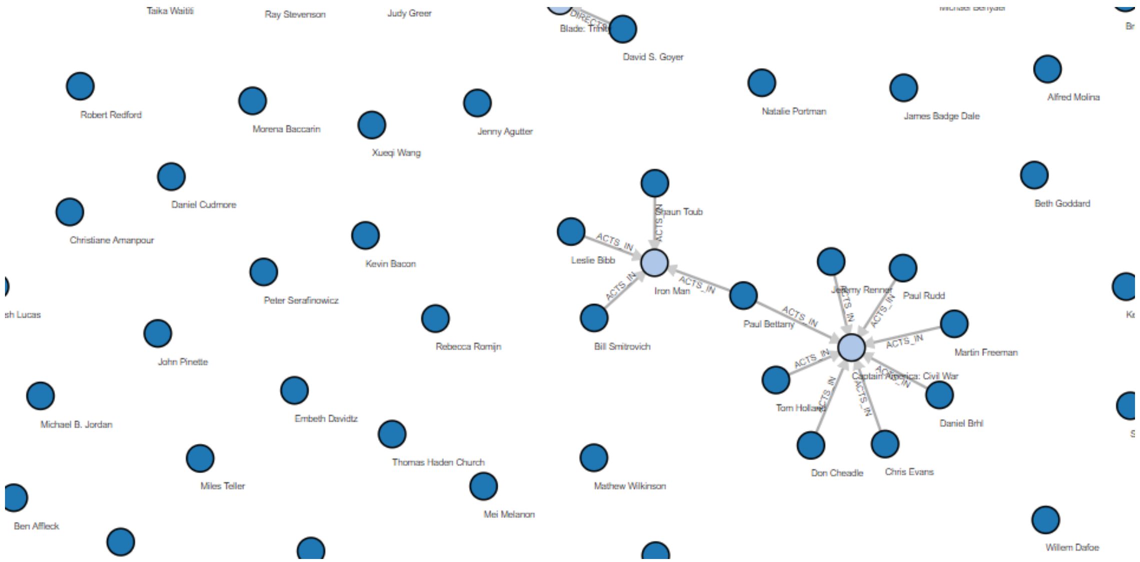
Hands On

1.4 Gremlin Language - Traversal

```
gremlin> g.V().hasLabel("movie").count()
==>45

gremlin> g.V().has('type','person').count()
==>502

gremlin> g.V().has('name','Thor').valueMap(true)
==>{id=245776, year=[2011], name=[Thor], rating=[7.0], runtime=[115], type=[movie], label=vertex}
```



http://localhost/graphexp/graphexp.html

Hands On

1.4 Gremlin Language - Traversal

Number of movies grouped by year?

```
g.V().has('type','movie').groupCount().by('year')
```

All the Thor movie titles?

```
g.V().has('name',textContains("Thor")).values("name")
```

Which of the movies are from the nineties?

```
g.V().hasLabel('movie').and(has('year',gt(1990)),has('year',lt(2000))).values("name")
```

Which people are both actors and directors?

```
g.V().hasLabel("person").and(where(out("DIRECTS")),where(out("ACTS_IN"))).values("name")
```

Which is the movie with the highest rating?

```
g.V().hasLabel("movie").order().by("rating",Order.decr).limit(1).values("name")
```

How many minutes is the runtime average?

```
g.V().hasLabel("movie").values("runtime").mean()
```

http://tinkerpop.apache.org/docs/current/reference/#graph-traversal-steps
http://tinkerpop.apache.org/javadocs/3.3.1/core/org/apache/tinkerpop/gremlin/process/traversal/dsl/graph/GraphTraversal.html
http://docs.janusgraph.org/0.2.0/search-predicates.html
http://sql2gremlin.com/

Hands On

1.4 Gremlin Language - Traversal

How many movies have more than two hours?
g.V().hasLabel('movie').has('runtime',gt(120)).count()

Who is the director of Blade?
g.V().has("name","Blade").in('DIRECTS').values("name")

Which are the movies that have more than twenty actors?
g.V().hasLabel("movie").where(inE('ACTS_IN').count().is(gt(20))).values("name")

Top 3 movies by number of actors?
g.V().hasLabel("movie").order().by(inE('ACTS_IN').count(),Order.decr).limit(3).values("name")

Which actors played “The Hulk”?
g.E().has("character","The Hulk").outV().dedup().values("name")

http://tinkerpop.apache.org/docs/current/reference/#graph-traversal-steps
http://tinkerpop.apache.org/javadocs/3.3.1/core/org/apache/tinkerpop/gremlin/process/traversal/dsl/graph/GraphTraversal.html
http://docs.janusgraph.org/0.2.0/search-predicates.html
http://sql2gremlin.com/

Hands On

1.5 Gremlin Language - Mutating

Add new movie node for the movie Captain Marvel (http://www.imdb.com/title/tt4154664/)
Add new person nodes for its directors
Add new person nodes for its actors
Create the needed edges between them.

NOTE: Use gremlin to check if the Actors/Director already exist in the graph before adding.

g.addV(*label*).property(*prop name*,*prop value*)
g.V().has(*prop name*,*prop value*).addE(*label*).to(g.V().has(*prop name*,*prop value*))

http://tinkerpop.apache.org/docs/current/reference/#traversal

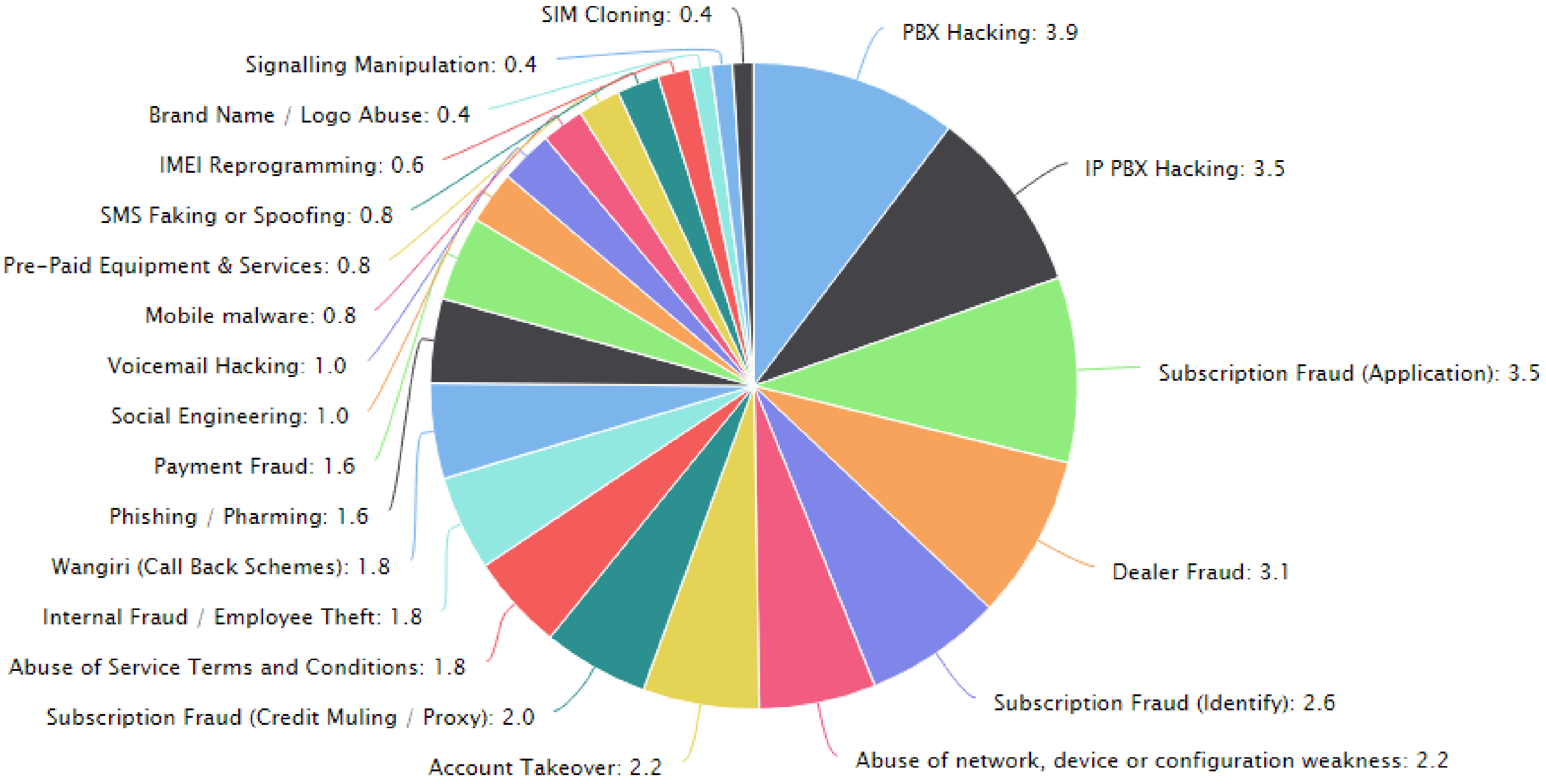
Hands On

2.1 THE IMPACT OF FRAUD

Fraud amounts to **$38.1 billion** annually representing **1.69%** of all Telecom revenues

(based on estimations from CFCA of 2015)

- **Fraudsters are everywhere** and Operators are always desirable targets
- The frequency and sophistication of **fraudulent activity on networks is rising**
- The wider business scope of Operators has **multiplied the areas where fraud can occur**
- **Black-box systems do not adapt** well to this new reality



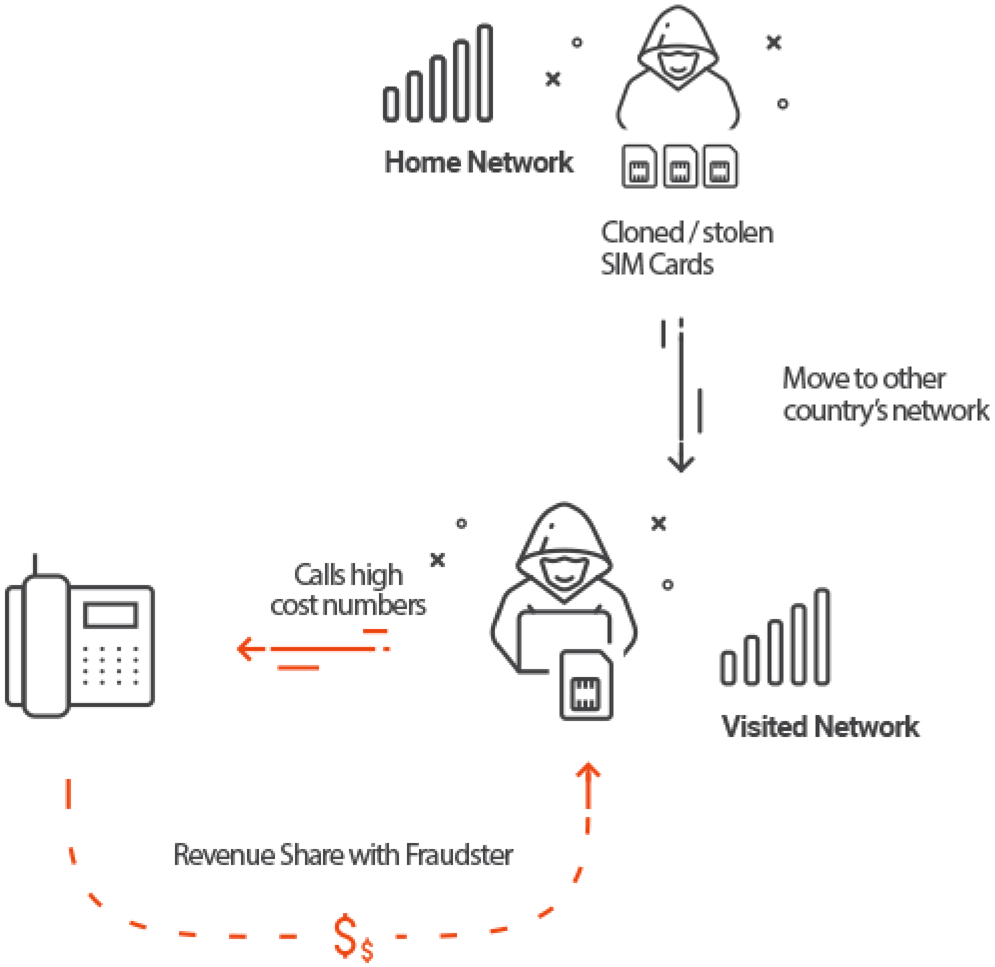
CFCA 2015 Survey - Fraud Losses by Method in $ USD Billions

Hands On

2.2 Detecting Fraud

High Usage to Risky Destinations/Numbers;

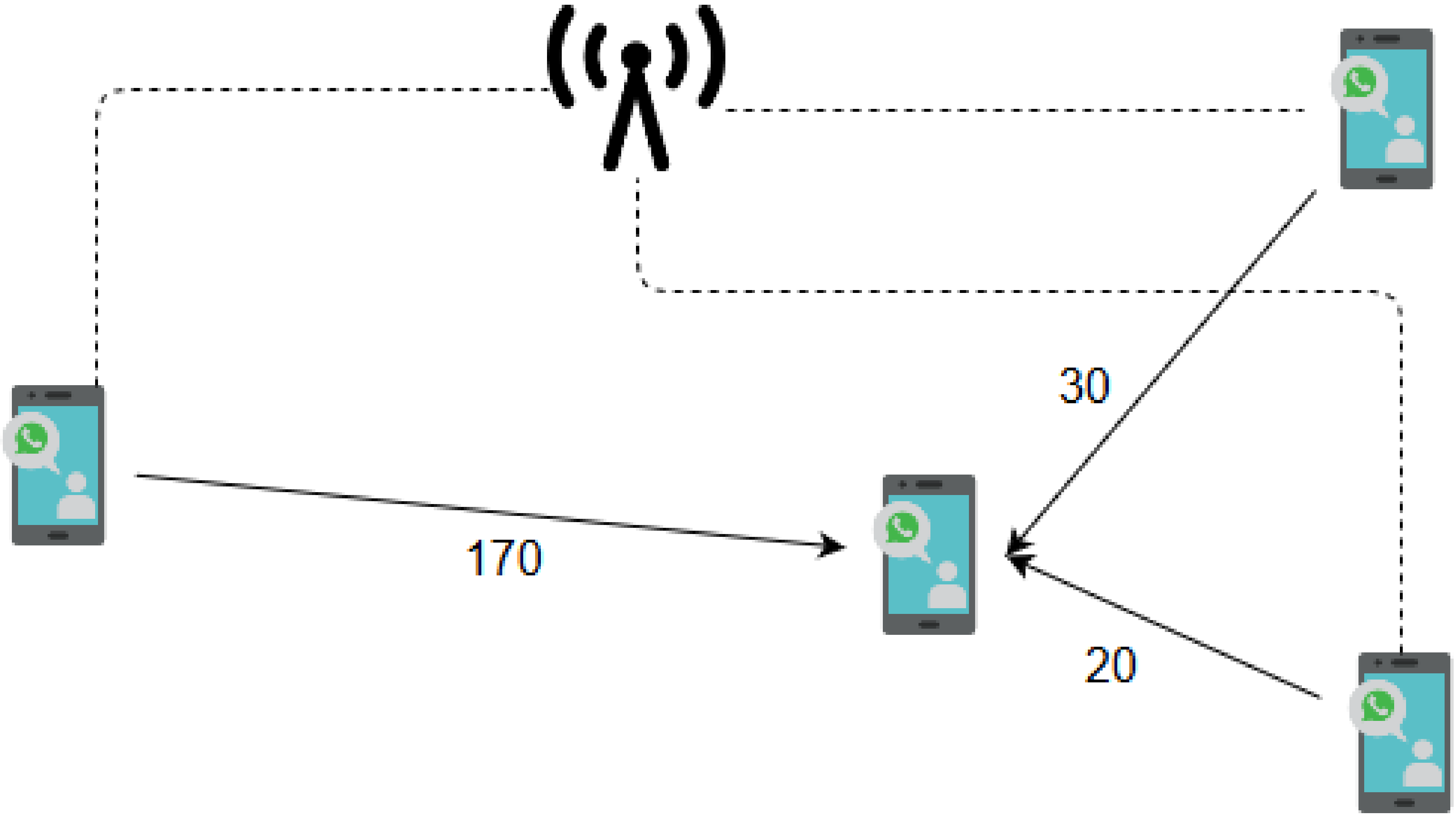
International High Usage: audit the traffic, based in the CDRs generated at the network switches, to identify and alarm the scenarios where the traffic to international hot-listed numbers is greater than a define threshold.


Home Network
Cloned / stolen SIM Cards
Move to other country's network
Calls high cost numbers
Visited Network
Revenue Share with Fraudster
$$

Hands On

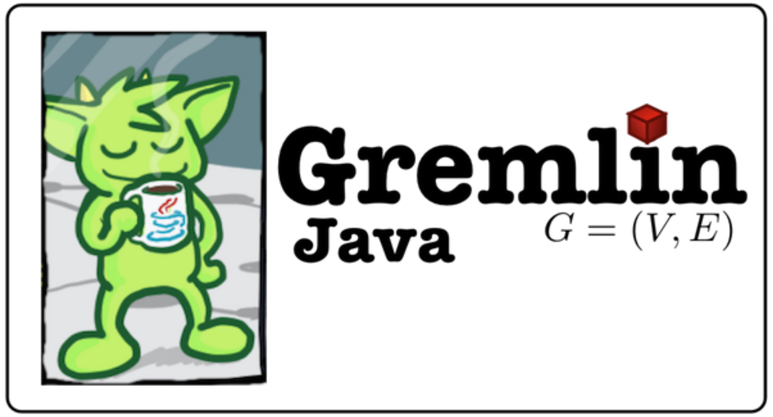
2.2 Detecting Fraud

High Usage to Risky Destinations/Numbers;



Hands On

2.3 Java



1. Open ExtractData.java
2. Go to http://localhost/graphexp/graphexp.html
3. Choose advanced mode and use the query:

```
nodes = graph.traversal().V()
edges = graph.traversal().E()
[nodes.toList(), edges.toList()]
```

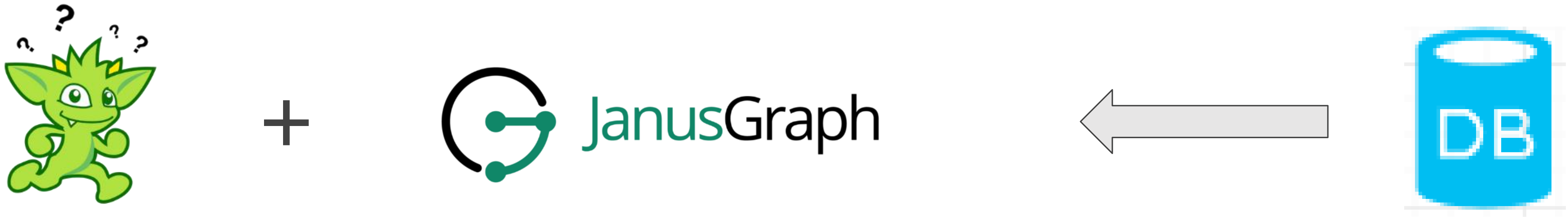
https://javadoc.io/doc/org.janusgraph/janusgraph-core/0.2.0

Hands On

2.3 From Database to Graph

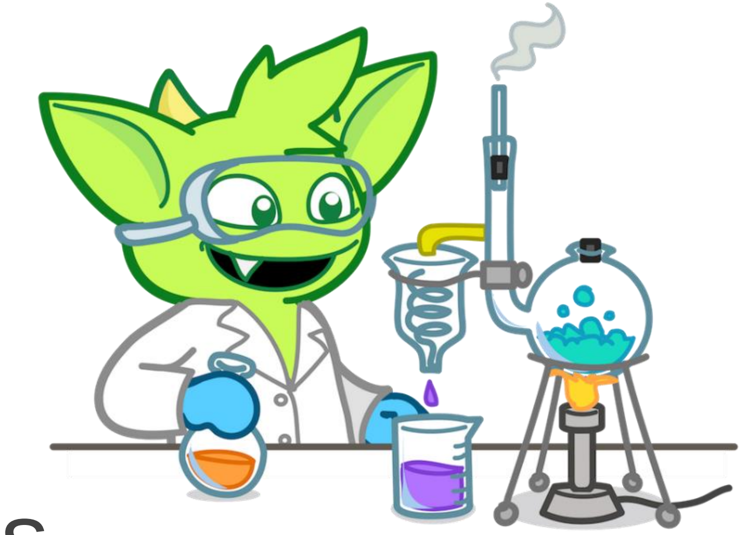
We found three options(till now)

- Generate the graphml file
- Generate and execute gremlin commands
- Execute Gremlin commands with remote



Conclusion

Highlights

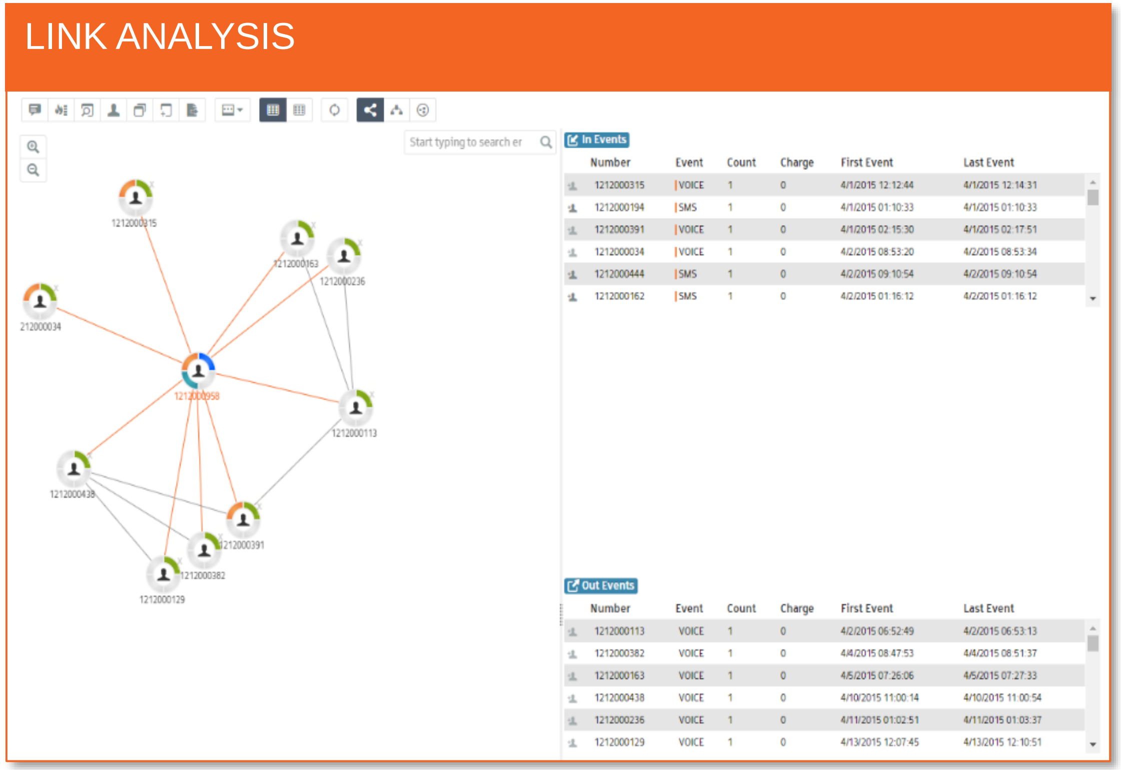


- Data modeling is very important.
- Is not that hard, but it takes time.
- It can be used in real world applications.
- If you choose one vendor that supports **Apache TinkerPop**, your are good to go.

Conclusion

Next Steps

- Improve the data loading process
- Use storage backend for persistence
- Play with Indexes
- Develop a user interface
- Evaluate some of the paid options



In Events

Number	Event	Count	Charge	First Event	Last Event
1212000315	VOICE	1	0	4/1/2015 12:12:44	4/1/2015 12:14:31
1212000194	SMS	1	0	4/1/2015 01:10:33	4/1/2015 01:10:33
1212000391	VOICE	1	0	4/1/2015 02:15:30	4/1/2015 02:17:51
1212000034	VOICE	1	0	4/2/2015 08:53:20	4/2/2015 08:53:34
1212000444	SMS	1	0	4/2/2015 09:10:54	4/2/2015 09:10:54
1212000162	SMS	1	0	4/2/2015 01:16:12	4/2/2015 01:16:12

Out Events

Number	Event	Count	Charge	First Event	Last Event
1212000113	VOICE	1	0	4/2/2015 06:52:49	4/2/2015 06:53:13
1212000382	VOICE	1	0	4/4/2015 08:47:53	4/4/2015 08:51:37
1212000163	VOICE	1	0	4/5/2015 07:26:06	4/5/2015 07:27:33
1212000438	VOICE	1	0	4/10/2015 11:00:14	4/10/2015 11:00:54
1212000236	VOICE	1	0	4/11/2015 01:02:51	4/11/2015 01:03:37
1212000129	VOICE	1	0	4/13/2015 12:07:45	4/13/2015 12:10:51

References:

https://medium.com/basecs/a-gentle-introduction-to-graph-theory-77969829ead8

http://tinkerpop.apache.org/

http://kelvinlawrence.net/book/Gremlin-Graph-Guide.html#gs

http://janusgraph.org/

Conclusion

Questions?

Q&A

https://goo.gl/RQ3rFW

THANK YOU

PEDRO.MPIRES@WEDOTEHCNOLOGIES.COM

RICARDO.FARIA@WEDOTECHNOLOGIES.COM

http://www.wedotechnologies.com/en/careers/

Know the unknown