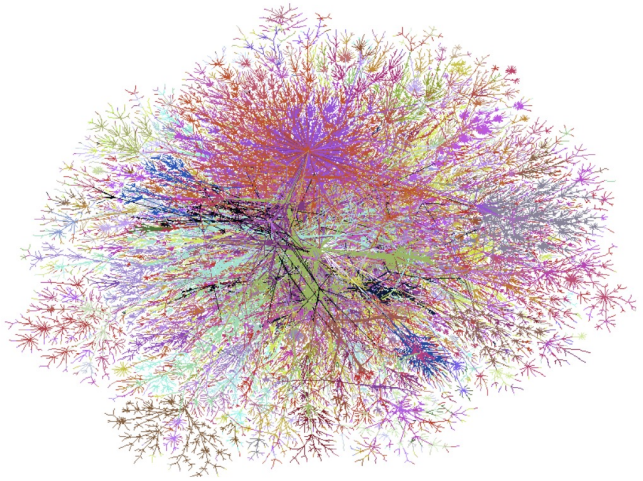
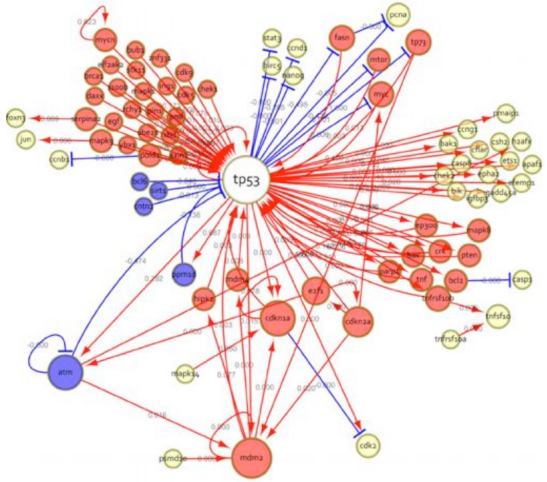
Apache GraphAr: An Open-Source Standard File Format for Graph Data Storage and Retrieval

Jingbo Xu
Institute for Intelligent Computing
Alibaba

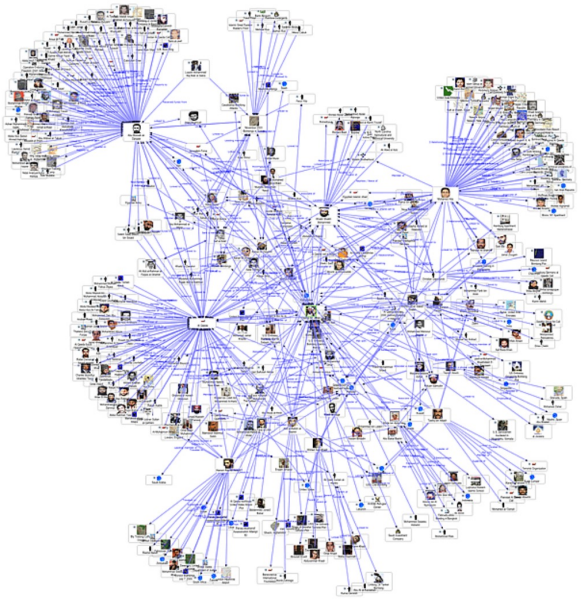
Graph Data is Everywhere



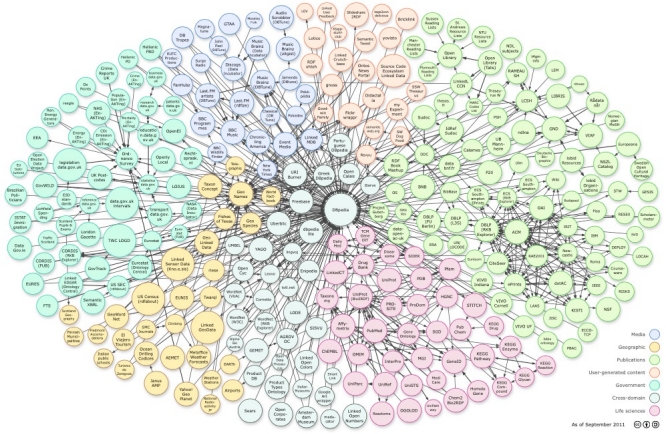
Web



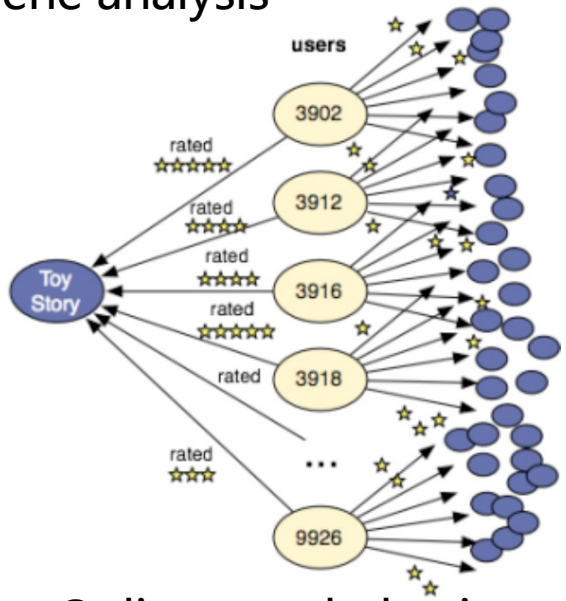
Gene analysis



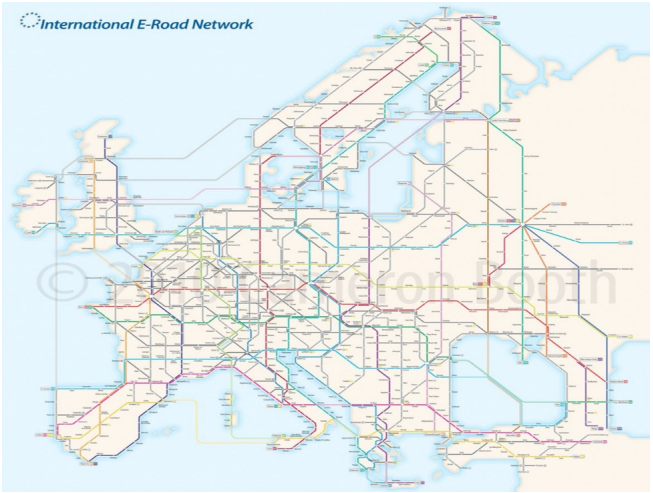
Social network



Knowledge base

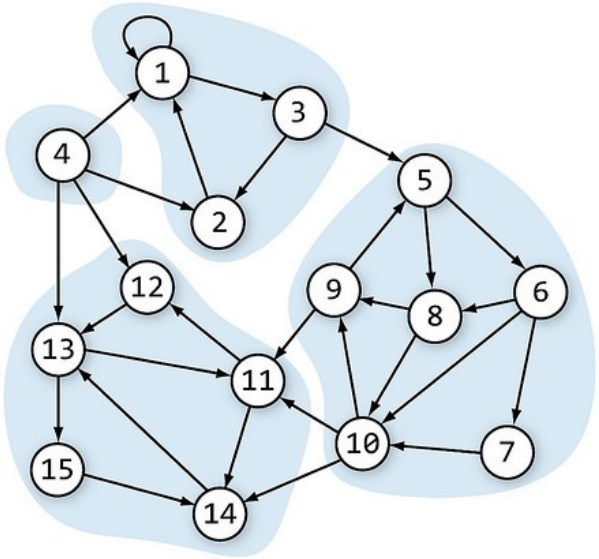


Online user behavior

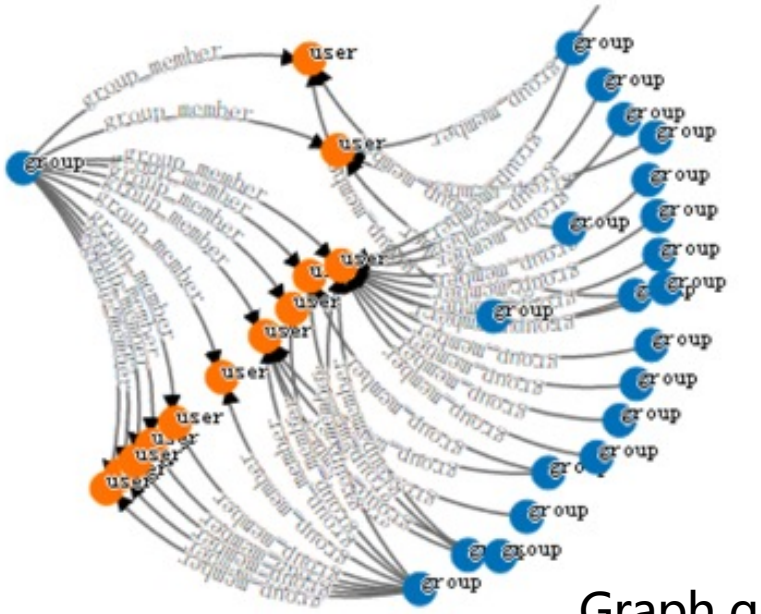


Traffic graph

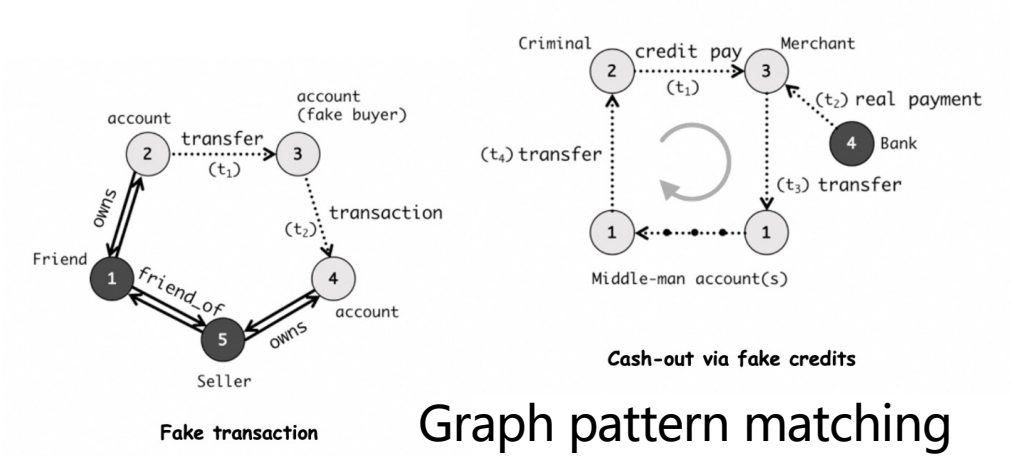
Diverse Graph Computing Workloads



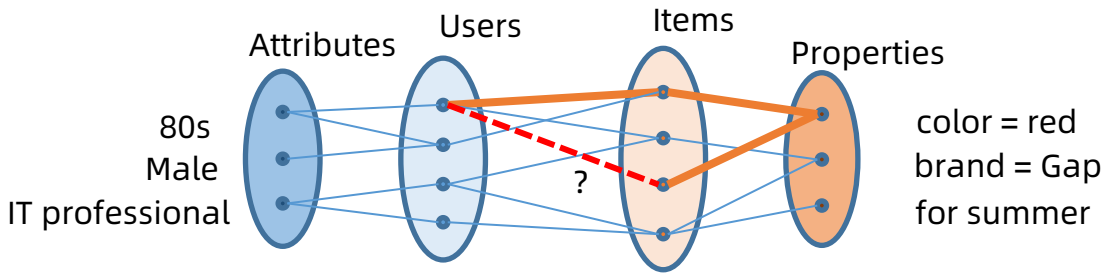
Iterative graph analytics, e.g., PageRank, LPA…



Graph querying/traversal



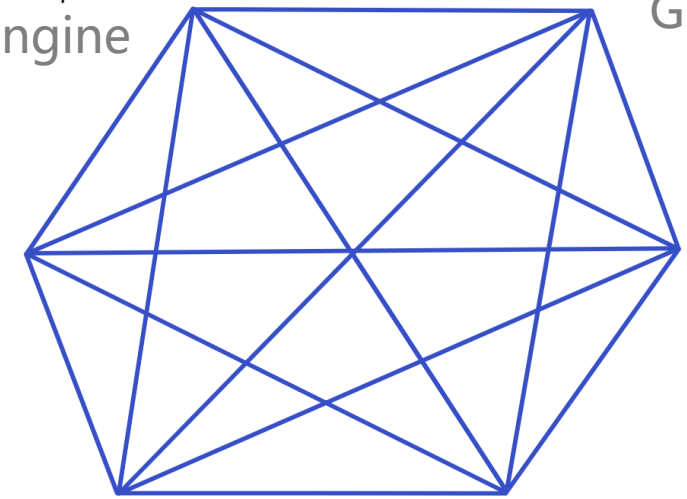
Graph pattern matching



Graph learning

Various Graph Systems

- Graph DBs, GNN Frameworks, Graph Engines, KBs….
- **How do they share graph data?**



APACHE GIRAPH

Spark GraphX
Computing Engine

TuGraph™
Graph Database

TigerGraph

GraphScope
Graph Computing Engine

neo4j
Graph Database

Amazon Neptune

Azure Cosmos DB
Graph Database

PyG
GNN Framework

JanusGraph

ArangoDB

What is GraphAr

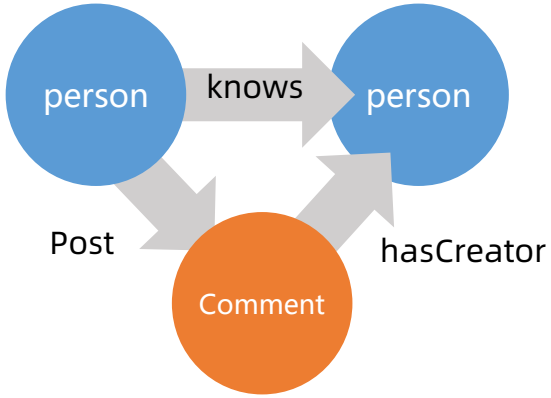
- Short for **Graph Ar**chive
- An open source, standard data file format for graph data storage & retrieval.
- Designed for
  - Serving as a standardized file format for importing, exporting and archiving of the graph data which can be used by diverse existing systems, reducing the overhead when various systems co-work.
  - As a direct data source for (out-of-core) graph processing applications.

Key Features

- **Apache incubating**, open-source and vendor neutral
- The file format supports the property graphs and different representations for the graph topology (COO, CSR and CSC).
- It is compatible with existing widely-used file formats including ORC, Parquet (and less ideally CSV).
- Apache Spark can be utilized to generate, load and transform GraphAr files.
- It is convenient for use in a variety of single-machine/distributed graph processing systems, databases, and other downstream computing tasks.
- It enables users to conveniently perform operations without modifying the payload files, such as appending new vertices, adding new properties, or constructing a new graph with a set of selected vertices and edges.

Design of GraphAr: A Running Example

- A property graph and its vertices



internal vid	id	firstName	lastName	gender
0	933	Mahinda	Perera	male
1	6597069767117	Eli	Peretz	female
2	10995116278700	Joseph	Anderson	female
...	...	...	...	...
903	32985348834100	Bruno	Oliveira	male

Logical Table of Vertices

internal vid	id
0	933
1	6597069767117
...	...
499	13194139534267

./vertex/person/id/chunk0

internal vid	firstName	lastName	gender
0	Mahinda	Perera	male
1	Eli	Peretz	female
...	...	...	...
499	Asha-Rose	Chung	male

./vertex/person/firstName_lastName_gender/chunk0

internal vid	id
500	15393162788965
501	15393162789614
...	...
903	32985348834100

./vertex/person/id/chunk1

internal vid	firstName	lastName	gender
500	Hans	Becker	male
501	Adi	Cohen	female
...	...	...	...
903	Bruno	Oliveira	male

./vertex/person/firstName_lastName_gender/chunk1

Physical Table of Vertices

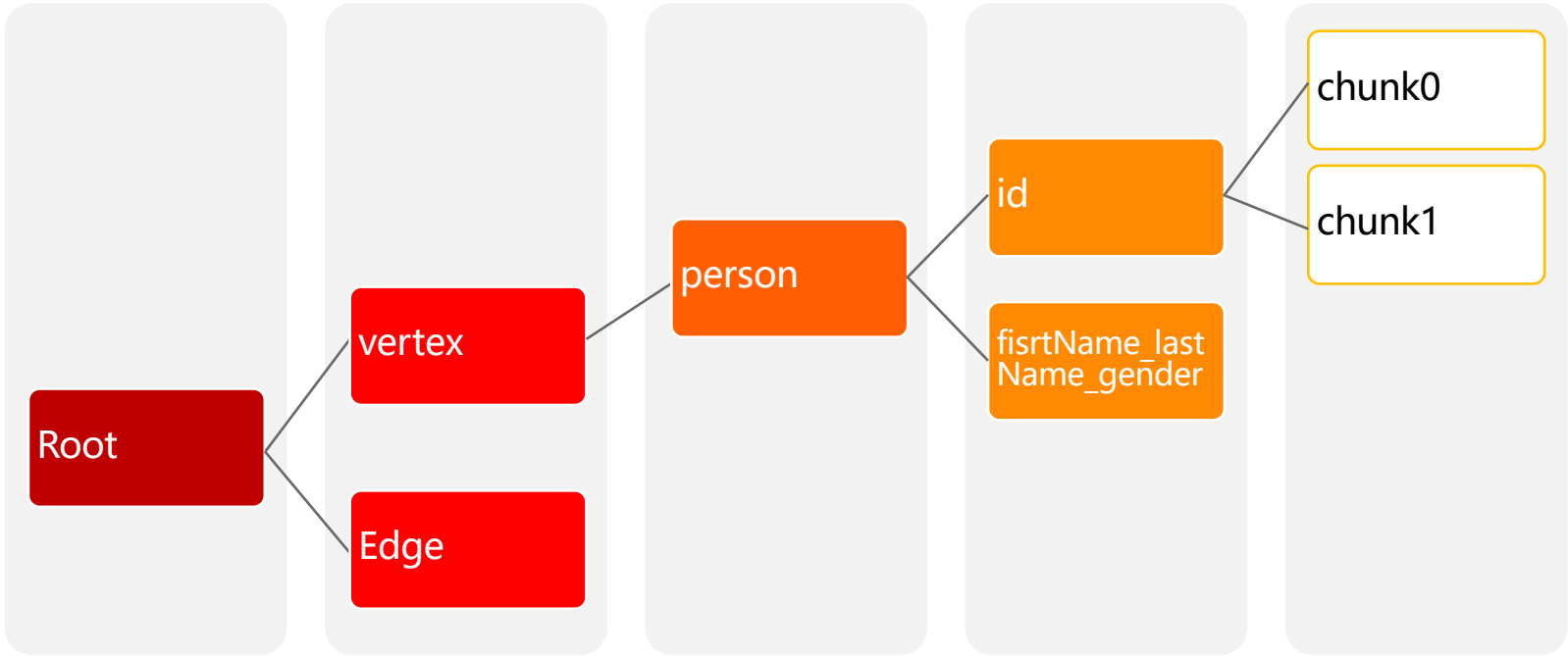
Chunk size = 500

Vertices in GraphAr: Metadata and Payload Files

Metadata person.vertex.yml

```yaml
label: person
chunk_size: 100
prefix: vertex/person/
property_groups:
  - properties:
      - name: id
        data_type: int64
        is_primary: true
    prefix: id/
    file_type: csv
  - properties:
      - name: firstName
        data_type: string
        is_primary: false
      - name: lastName
        data_type: string
        is_primary: false
      - name: gender
        data_type: string
        is_primary: false
    prefix: firstName_lastName_gender/
    file_type: csv
version: gar/v1
```

Payload files



Edges in GraphAr

Logical Table of Edges

internal vid	offset
0	0
1	3
...	...
901	6625
902	6626
903	6626
	6626

source	destination	creationDate
0	87	2010-07-30T15:19:53.298+0000
0	849	2010-10-27T02:33:06.288+0000
0	623	2012-04-12T13:56:58.931+0000
1	538	2011-03-25T14:49:23.134+0000
1	539	2011-06-16T11:28:39.623+0000
...	...	...
1	58	2011-06-10T17:47:19.432+0000
...	...	...
900	826	2012-08-21T00:02:08.530+0000
901	252	2012-08-13T10:11:20.606+0000

Physical Table of Edges

edges of vertex chunk 0

internal vid	offset
0	0
1	3
...	...
499	3772
	3778

./edge/person_knows_person/ordered_by_source/offset/chunk0

source	destination
0	87
...	...
164	829

./edge/person_knows_person/ordered_by_source/adj_list/part0/chunk0

creationDate
2012-04-21T19:08:41.647+0000
...
2012-08-10T02:49:19.288+0000

./edge/person_knows_person/ordered_by_source/creationDate/part0/chunk0

source	destination
164	30
...	...
269	565

./edge/person_knows_person/ordered_by_source/adj_list/part0/chunk1

creationDate
2012-06-26T02:41:08.148+0000
...
2012-03-10T06:07:41.141+0000

./edge/person_knows_person/ordered_by_source/creationDate/part0/chunk1

source	destination
269	321
...	...
499	628

./edge/person_knows_person/ordered_by_source/adj_list/part0/chunk2

creationDate
2012-02-19T06:42:02.399+0000
...
2012-08-13T10:11:20.606+0000

./edge/person_knows_person/ordered_by_source/creationDate/part0/chunk2

edges of vertex chunk 1

internal vid	offset
500	0
501	1
...	...
903	2848
	2848

./edge/person_knows_person/ordered_by_source/offset/chunk1

source	destination
500	623
...	...
637	704

./edge/person_knows_person/ordered_by_source/adj_list/part1/chunk0

creationDate
2010-07-30T15:19:53.298+0000
...
2010-06-11T19:23:42.146+0000

./edge/person_knows_person/ordered_by_source/creationDate/part1/chunk0

source	destination
638	375
...	...
793	884

./edge/person_knows_person/ordered_by_source/adj_list/part1/chunk1

creationDate
2010-05-16T17:41:47.623+0000
...
2011-12-22T17:56:13.491+0000

./edge/person_knows_person/ordered_by_source/creationDate/part1/chunk1

source	destination
793	662
...	...
901	252

./edge/person_knows_person/ordered_by_source/adj_list/part1/chunk2

creationDate
2012-01-04T13:29:11.784+0000
...
2012-08-03T01:00:51.312+0000

./edge/person_knows_person/ordered_by_source/creationDate/part1/chunk2

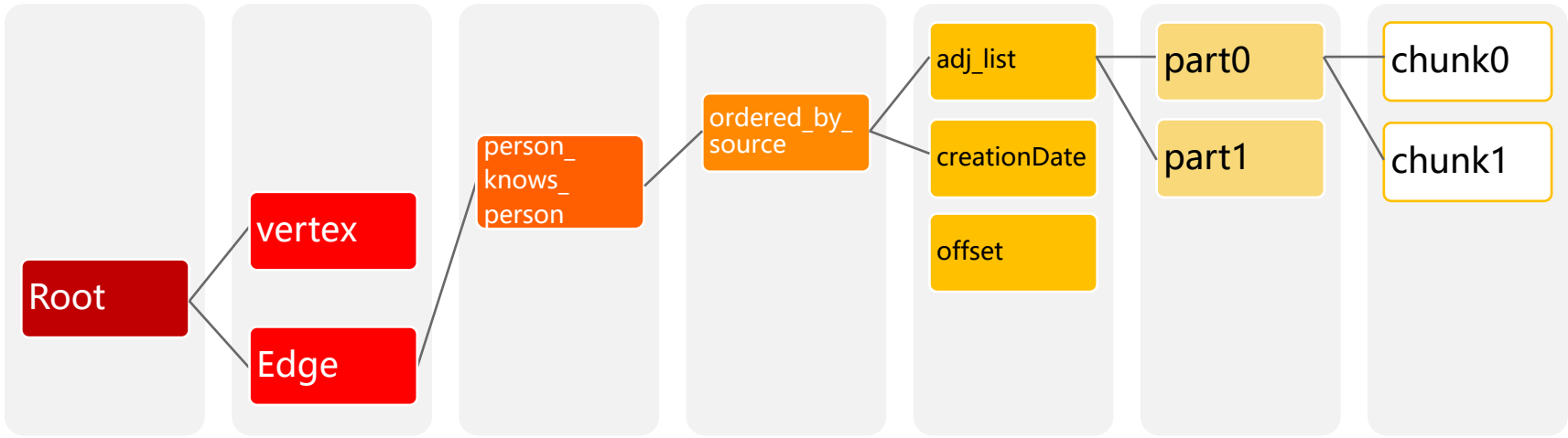
Edges in GraphAr: Metadata and Payload Files

Metadata

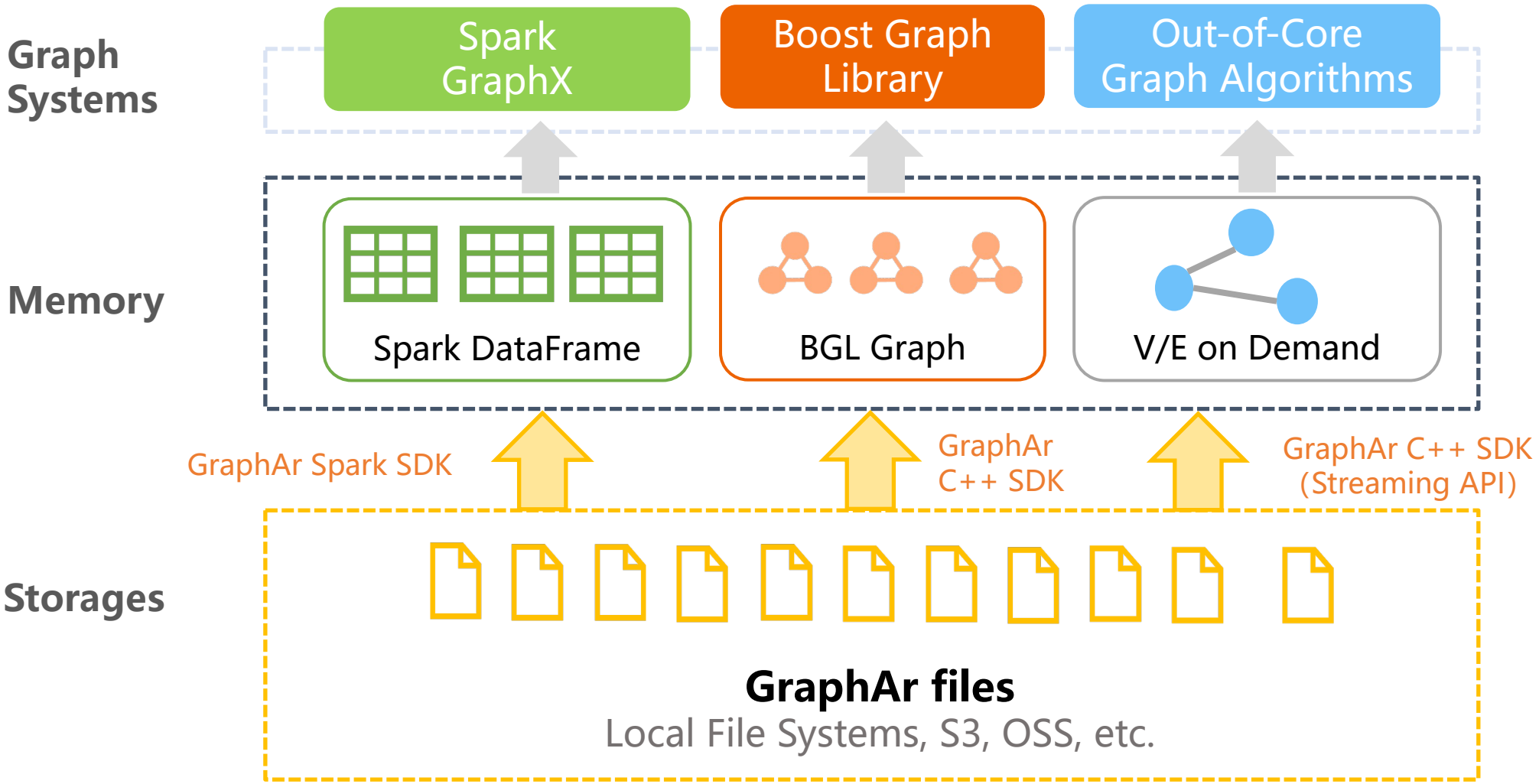
person_knoews_person.edge.yml

```yaml
src_label: person
edge_label: knows
dst_label: person
chunk_size: 1024
src_chunk_size: 100
dst_chunk_size: 100
directed: false
prefix: edge/person_knows_person/
adj_lists:
  - ordered: true
    aligned_by: src
    prefix: ordered_by_source/
    file_type: csv
    property_groups:
      - prefix: creationDate/
        file_type: csv
        properties:
          - name: creationDate
            data_type: string
            is_primary: false
  - ordered: true
    aligned_by: dst
    prefix: ordered_by_dest/
    file_type: csv
    property_groups:
      - prefix: creationDate/
        file_type: csv
        properties:
          - name: creationDate
            data_type: string
            is_primary: false
version: gar/v1
```

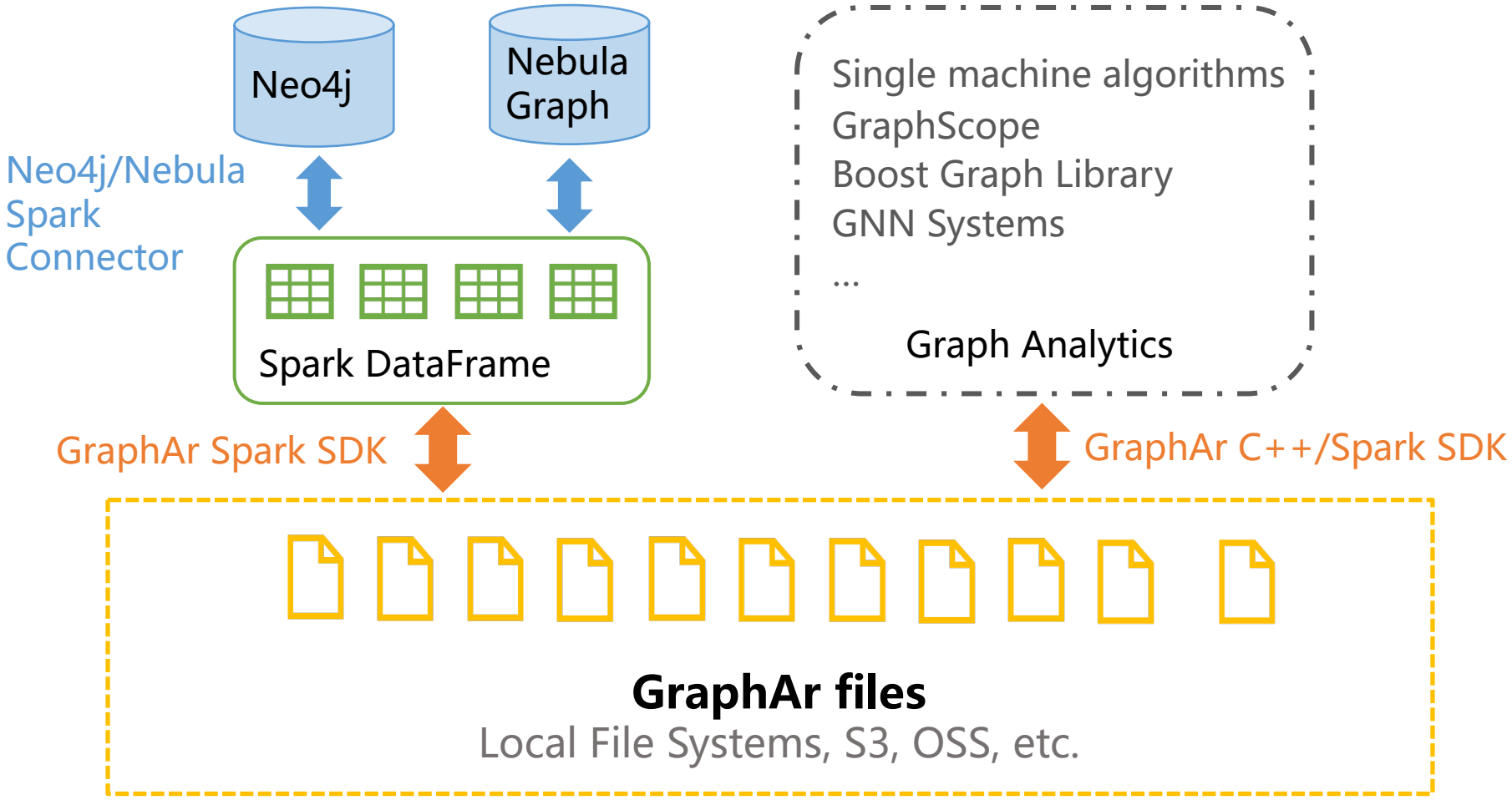
Payload files



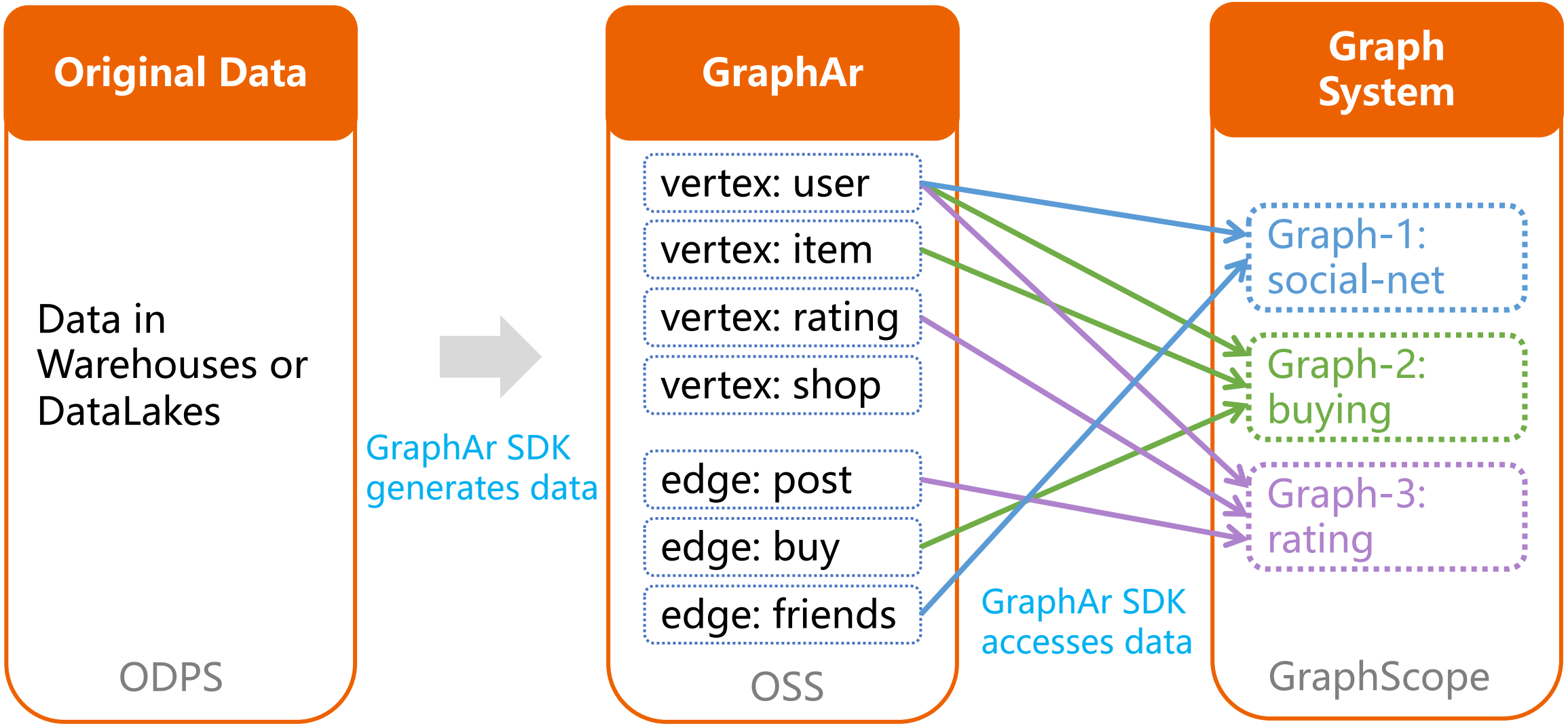
GraphAr in Action: As Graph Data Source



GraphAr in Action: As Archive of Graph DBs



GraphAr in Action: Graph Transformation



Future of GraphAr

- Support more file formats, more standard and user-defined data types.
- More graph features: RDF, time-series
- Encoding optimizations.
- Complete Spark suite to transform & create GraphAr files.
- Integrations with popular graph systems, such as Neo4j, Nebula, TuGraph, PyG …
- Explore the use of GraphAr for data lake of graphs.

Welcome to join forces with us!

Thank you!

Learn more at https://graphar.apache.org

Initiated and donated by GraphScope Team from Alibaba