Graph Analytics and Machine Learning in Spark

Data Application Lab

Graph Analytics

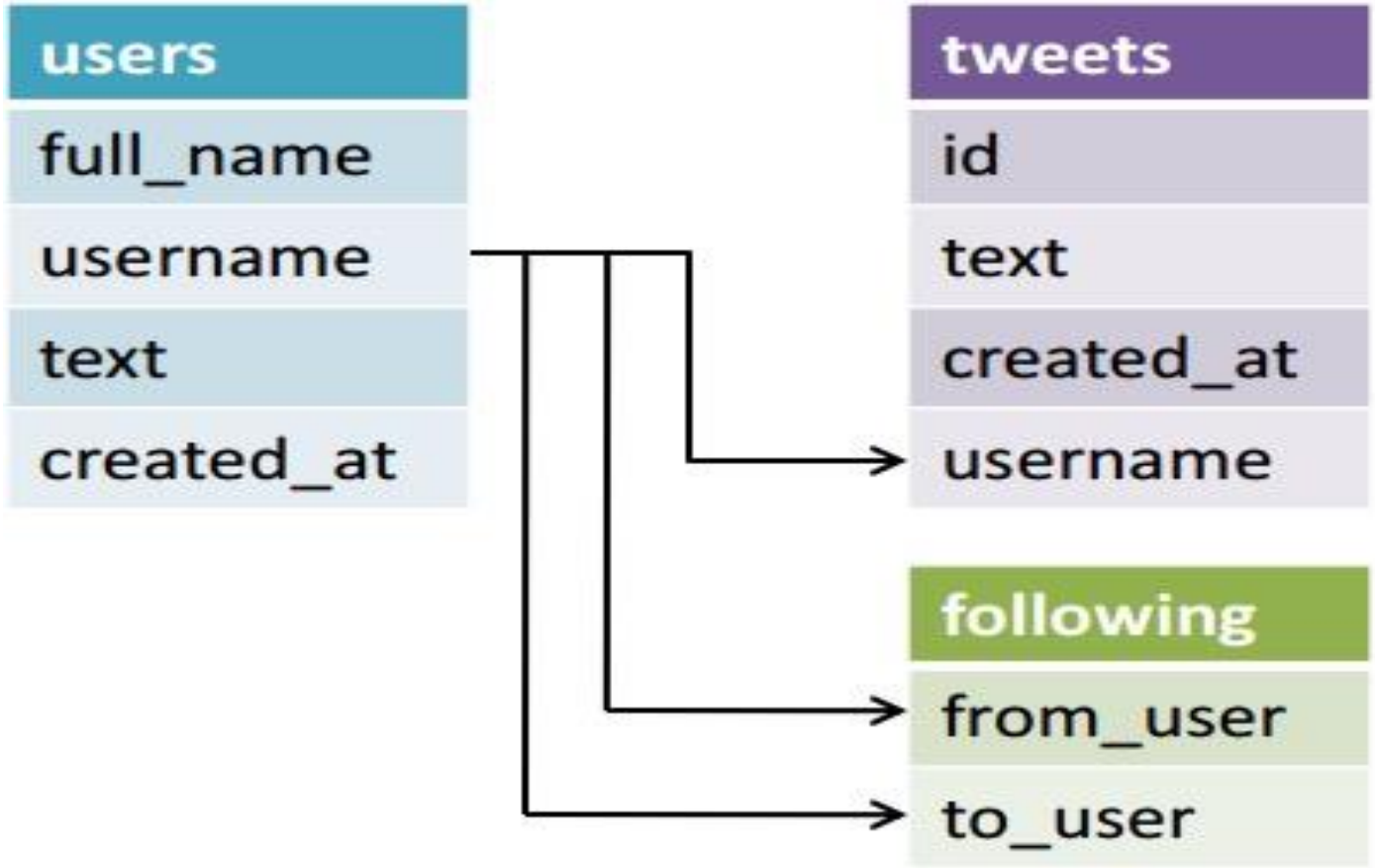
- What is graph database
- Applications
- Spark GraphX/GraphFrame
- Demo

Machine Learning

- Machine learning at scale
- Spark ML in action

Relational Database

Data Application Lab



- Schema
- Table
- Key
- Join
- View

No-SQL Database

Data Application Lab

- Key-Value
- Column based
- Document based
- Graph database

- Data model
- Data structure
- Scaling
- Development model

Data is more connected

Data Application Lab

- Text
- Hypertext
- RSS
- Blogs
- Comments
- Review
- Endorse
- Message
- Locations

What is Graph Database

Data Application Lab

1. A database with graph structure
2. Each node knows its adjacent nodes
3. As the number of nodes increases, the cost of a local computation remains the same

Why graph database

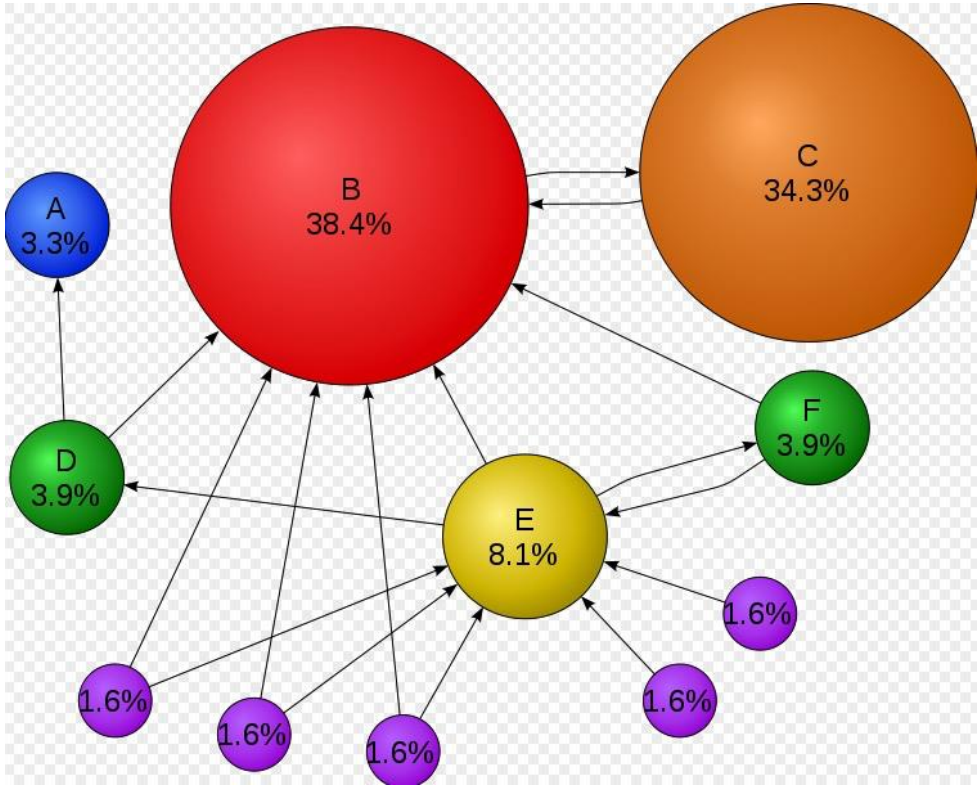
Data Application Lab

1. Data connected
2. Performance
3. Flexibility
4. Agility

PageRank Algorithm

Data Application Lab

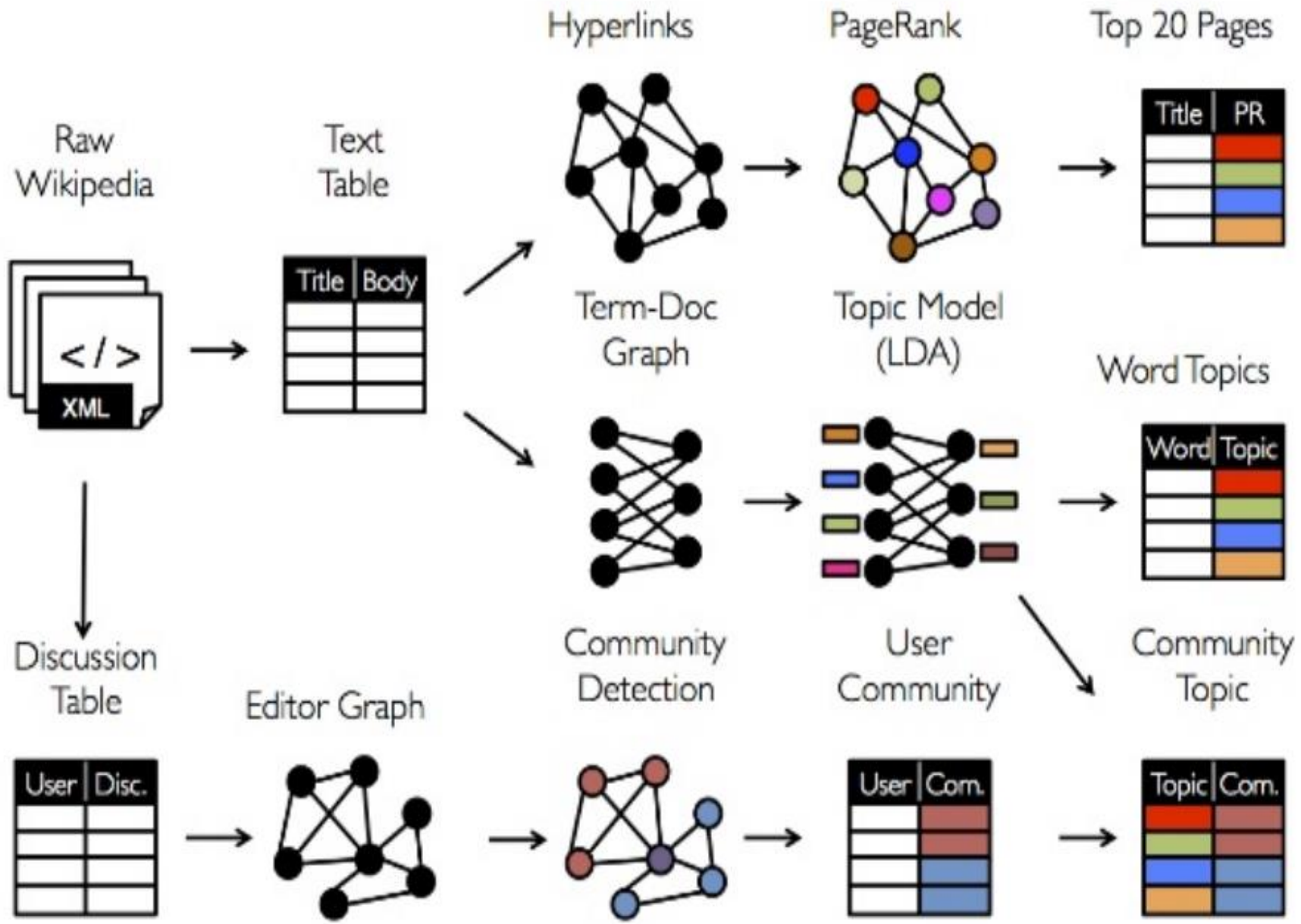
1. Rank page in Google search engine
2. Roughly estimate how important the website is
3. Count the number and quality of links at a page
4. Page C has a higher PageRank than Page E, even though there are fewer links to C; the one link to C comes from an important page and hence is of high value



Use Cases

Data Application Lab

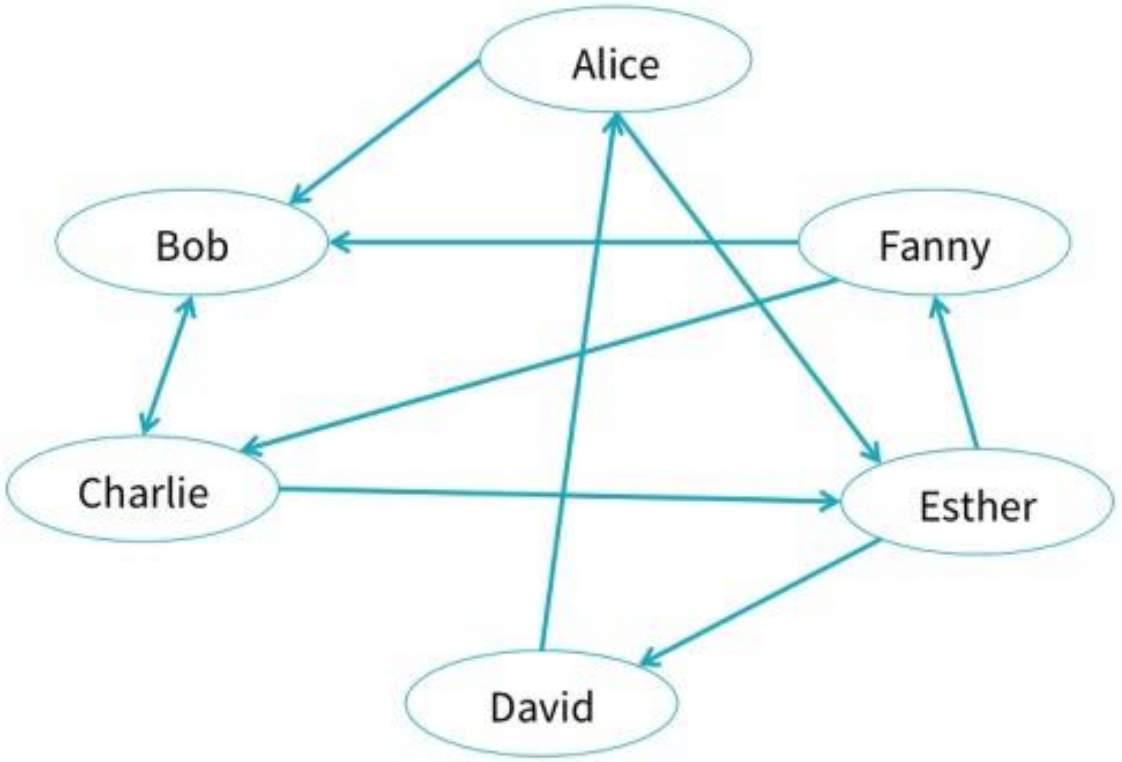
1. Fraud detection
2. Graph-based search
3. Network IT operations
4. Real-time recommendation
5. Social network
6. Identity and access management



Data Application Lab

GraphX - Demo

Data Application Lab



id	name	age
a	Alice	34
b	Bob	36
c	Charlie	30
d	David	29
e	Esther	32
f	Fanny	36

src	dst	relationship
a	e	friend
f	b	follow
c	e	friend
a	b	friend
b	c	follow
c	b	follow
f	c	follow
e	f	follow
e	d	friend
d	a	friend